# TOWARDS CLASSIFICATION PARITY ACROSS COHORTS

**Aarsh Patel**
Department of Computer Science
University of Massach
Pittsburgh, PA 15213, USA
aarshpatel@umass.edu

**Rahul Gupta**
Amazon Alexa
Cambridge, Massachusetts
gupra@amazon.com

**Mukund Harakere**
Amazon Alexa
Cambridge, Massachusetts
harakere@amazon.com

**Satyapriya Krishna**
Amazon Alexa
Cambridge, Massachusetts
satyapk@amazon.com

**Aman Alok**
Amazon Alexa
Cambridge, Massachusetts
alokaman@amazon.com

**Peng Liu**
Amazon Alexa
Cambridge, Massachusetts
liupng@amazon.com

## ABSTRACT

Recently, there has been a lot of interest in ensuring algorithmic fairness in machine learning where the central question is how to prevent sensitive information (e.g. knowledge about the ethnic group of an individual) from adding 'unfair' bias to a learning algorithm (Feldman et al. (2015), Zemel et al. (2013)). This has led to several debiasing algorithms on word embeddings (Qian et al. (2019) , Bolukbasi et al. (2016)), coreference resolution (Zhao et al. (2018a)), semantic role labeling (Zhao et al. (2017)), etc. Most of these existing work deals with explicit sensitive features such as gender, occupations or race which doesn't work with data where such features are not captured due to privacy concerns. In this research work, we aim to achieve classification parity across explicit as well as implicit sensitive features. We define explicit cohorts as groups of people based on explicit sensitive attributes provided in the data (age, gender, race) whereas implicit cohorts are defined as groups of people with similar language usage. We obtain implicit cohorts by clustering embeddings of each individual trained on the language generated by them using a language model. We achieve two primary objectives in this work : [1.] We experimented and discovered classification performance differences across cohorts based on implicit and explicit features , [2] We improved classification parity by introducing modification to the loss function aimed to minimize the range of model performances across cohorts.

## 1 INTRODUCTION

Machine learning has proven to be useful in several applications such as speech recognition (Graves et al. (2013)), image classification (Krizhevsky et al. (2012)), reading comprehension (Hermann et al. (2015)), etc and growing in other critical application such as healthcare (Esteva et al. (2019)). With such rapid increase in data driven solutions making important decisions in our lives, it is worth discussing the possibility of bias in these models. Apart from the generally discussed algorithmic bias, there are several other types of bias such as, representational bias (Suresh & Guttag (2019)) due to the lack of diversity in data samples, historical bias (Suresh & Guttag (2019)) caused due to the existing societal biases unknowingly making its way to the data generation process, etc that can negatively impact decision making.

Recently, there has been a lot of work in removing biases through debiasing techniques motivated by algorithmic fairness (Gonen & Goldberg (2019) Zhao et al. (2018b)). All these models works only on explicit attributes such as gender or occupation which comes with two downsides. First, explicit attributes may not be available in a generic dataset due to privacy concerns. Secondly, gender might not be the only attribute causing unfair bias in the model. Hence, we experimented with explicit (gender) as well as implicit (hidden) sensitive attributes and improved classification parity for both of them.

# 2 PROPOSED METHOD

## 2.1 BACKGROUND

In the last few years, several fairness definitions (Hardt et al. (2016)Dwork et al. (2012)Kusner et al. (2017)Berk et al. (2018)) have been proposed taking different viewpoints under consideration. In this research work, we worked only on group level fairness definitions for two reasons. First, we are aiming to provide classification parity across cohorts before going down to individual performance parity. Second, its impossible to satisfy all the fairness definitions at once (Kleinberg et al. (2016)) without complex constraints, hence, we experimented with group level fairness for simplicity, i.e, demographic parity (Verma & Rubin (2018)) and equalized odds(Hardt et al. (2016)). There are multiple ways of implementing these definitions to debias our model and reach classification parity. In our work, we try to implement them by making modification to the loss function only and avoid any complex model architecture changes. This helps in scaling our solution to different applications.

## 2.2 OBJECTIVE

With the goal to achieve classification parity across multiple cohorts, we propose the optimization of loss function in equation 1, where $(\boldsymbol{x}, y)$ represents the pair of features and labels, respectively, in a dataset $\mathcal{D}$. $l(y, f(\boldsymbol{x}))$ is the loss incurred by a trainable function $f(\boldsymbol{x})$. $\mathcal{D}_i$ represents the cohort $i$ in the dataset and $\boldsymbol{x}_i, y_i$ are the datapoints belonging to the cohort $i$. The loss function has two components as depicted in the equation: the first component is an overall loss over the entire dataset and the second component enforces parity across cohorts. $\lambda$ is the weightage given to parity component when compared to the loss over the entire dataset.

$$\mathcal{L} = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} l(y, f(\boldsymbol{x})) + \lambda max_{i,j} \Big| \sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_i} l(y_i, f(\boldsymbol{x}_i)) - \sum_{(\boldsymbol{x}_j,y_j)\in\mathcal{D}_j} l(y_j, f(\boldsymbol{x}_j)) \Big| \tag{1}$$

We chose to minimize the difference between best and worst performing cohorts for the sake of simplicity. We also considered an alternate strategy to evaluate every pair during optimization. However, this strategy is expensive and determining weights for each pairwise difference is non-trivial. Additionally, this loss formulation allows the cohorts $i, j$ used in the optimization to either be coarse level cohorts (e.g. gender) or fine grained cohorts (e.g. a specific gender belonging to a specific ethnic background and a specific country). We use this inherent flexibility of the loss function to dynamically modify the granularity of cohorts during optimization. We use stochastic gradient descent to minimize $\mathcal{L}$, and we note that during each iteration, one may obtain a different cohort pair $i, j$ in the parity loss component.

## 2.3 COHORT DEFINITION

In the datasets of our interest, we use two broad categories of cohorts - explicit and implicit. We provide a brief description of these below.

### 2.3.1 EXPLICIT COHORTS

We define cohorts obtained based on attributes such as gender and location as explicit cohorts. These attributes could be directly observed for each individual in the dataset.

### 2.3.2 IMPLICIT COHORTS

We realize that available explicit attributes may not capture all the characteristics of a given individual. For instance, individuals may have not reported their ethnic background and/or some attributes may have not been recorded during data collection. To address this, we also determine implicit cohorts based on an individual's language usage. We obtain vector representations for each individual based on the language generated by them using a language model. In summary, we add an individual-id as additional token during language modeling task and use the *word-embedding* obtained for the individual-id token as their vector representation. These vector representations are clustered to define implicit cohorts.

# 3 EXPERIMENT

Our experiment followed a simple framework where we started with performance comparison across cohorts (implicit and explicit) and then retrained models with our proposed changes to the loss function. We measured classification parity in terms of the standard deviation of performance across cohorts.

## 3.1 DATASETS

We used three datasets for our experiments and their details are listed below.

| Dataset | Description | Classification Task |
|---|---|---|
| **Yelp Dataset (YD)** Yelp (2010) | Contains reviews from different businesses with a 5 star rating attached to every review. We added gender component in this dataset by using a publicly available tool (Genderize.io) which predicts gender by the name of the individual | Review rating prediction task |
| **TrustPilot Dataset (TPD)** Hovy et al. (2015) | Contains reviews from different business across the US and UK with a 5 star rating along with user metadata such as gender and location. | Review rating prediction task |
| **Internal Dataset (ID)** | Contains user utterances with labeled domain information (23 domain categories) along with metadata such as the gender, NLU(Natural Language Understanding) Score and ASR (Automatic Speech Recognition) score, which are the confidence scores of speech recognition and language understanding modules of the system | Domain Classification |

Table 1: Data description with corresponding classification task

## 3.2 DISPARITY IN MODEL PERFORMANCE

In the first segment of our experiments, we computed model performance across different cohorts for the corresponding tasks mentioned in Table 1 for each dataset.

### 3.2.1 EXPLICIT COHORTS

Explicit cohorts are based on explicit sensitive attributes such as gender or location. In the case of internal dataset, we used another set of explicit attributes, i.e, ASR (Automatic Speech Recognition) and NLU (Natural Language Understanding) scores since it represents the nativity of cohorts, i.e, people speaking a specific dialect will have similar ASR scores. We create two cohorts for each of these scores based on a predefined threshold($t$) and called it "NLU High" cohort with samples having NLU scores higher that $t$ and "NLU Low" cohort otherwise, similarly for ASR score. We observed disparities across all of the explicit cohorts and listed in table 2

### 3.2.2 IMPLICIT COHORTS

Implicit cohorts are based on implicit sensitive features which could be in terms of the linguistic differences based on regions (Labov et al. (1997)). We extracted such implicit cohort by training a language model for individuals based on their historical utterances, more details in Appendix B. We used k-means to cluster the embeddings, used k = 4 for the experiments. The performance difference across the cohorts are listed in table 3.

| Dataset | Sensitive Attribute | | Accuracy |
|---|---|---|---|
| **YD** | Gender | Male | 0.6306 |
| | | Female | 0.665 |
| **TPD** | Location (country) | UK Cohort | 0.8612 |
| | | US Cohort | 0.8955 |
| | Gender | Male | 0.8787 |
| | | Female | 0.9109 |
| **ID** | Gender | Male | |
| | | Female | 0.038 |
| | ASR Score | ASR High | |
| | | ASR Low | 0.049 |
| | NLU Score | NLU High | |
| | | NLU Low | 0.0514 |

Table 2: Model performance over explicit cohorts. We report performance differences for internal dataset (ID) only. For instance, abs(Accuracy(Male Cohort) - Accuracy(Female Cohort)) = 0.038

| Dataset | Implicit Cohort-1 (IC 1) | Implicit Cohort-2 (IC 2) | Implicit Cohort-3 (IC 3) | Implicit Cohort-4 (IC 4) |
|---|---|---|---|---|
| **YD** | 0.674 | 0.6511 | 0.6445 | 0.658 |
| **TPD** | 0.885 | 0.832 | 0.8901 | 0.8402 |
| **ID** | 0.0205 | | | |

Table 3: Model performance of implicit cohorts. We report std. deviation of accuracy on domain classification task for ID only.

### 3.3 Classification Parity

In order to reduce disparities discovered in the last section, we retrained our model with the proposed changes to the loss function and measured the disparity in terms of the standard deviation of model performances across cohorts over test set. We experimented with different values of $\lambda$ in equation 1 to observe the changes it causes to the parity and the overall model performance. We computed the standard deviation over a spectrum of cohorts by grouping multiple explicit or implicit sensitive attributes which gives a better picture about the robustness of our solution, more details in Appendix C.

| Dataset | No penalty | $\lambda = 0.5$ | $\lambda = 0.8$ |
|---|---|---|---|
| **ID** | 2.96 | 2.8 | 2.62 |
| **TPD** | 1.81 | 1.61 | 1.06 |
| **YD** | 1.55 | 1.42 | 1.38 |

Table 4: Standard deviation of model performance across cohorts for $\lambda = 0, 0.5, 0.8$

As shown in the table 4, we observe a drop in the standard deviation by 10-12 % with increased value of $\lambda$.

## 4 Conclusion

In this work, we experimented and found disparities in model performance across explicit and implicit cohorts. In order to fix these disparities, we added a penalty term in the loss function with the goal to minimize performance difference between the best and worst performing cohort. Our experiments showed that this change reduces disparity across explicit and implicit cohorts. We also

see improvements in parity as we increase the weight for this penalty term with minimal impact to the overall model performance for multiple cohorts.

## REFERENCES


Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533, 2018.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pp. 4349–4357, 2016.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24, 2019.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Genderize.io. Determine the gender of a name. URL `https://genderize.io/`.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. IEEE, 2013.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pp. 1693–1701, 2015.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, pp. 452–461. International World Wide Web Conferences Steering Committee, 2015.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

William Labov, Sharon Ash, and Charles Boberg. A national map of the regional dialects of american english.(telsur project.) department of linguistics, university of pennsylvania, 1997.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. Reducing gender bias in word-level language models with a gender-equalizing loss function. *arXiv preprint arXiv:1905.12801*, 2019.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7. IEEE, 2018.

Yelp. Yelp dataset challenge, 2010. URL `https://www.yelp.com/dataset/challenge`.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018a.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018b.

## A Fairness Definitions

- **Demographic Parity** : It ensures that the likelihood of an outcome is independent of whether the person belongs to any social group, known as sensitive attribute, for instance, gender is a sensitive attribute. (Verma & Rubin (2018))
- **Equalized Odds** : It requires to have the same rate for true positive as wells as false positives for all groups from sensitive attribute (eg. male and female). (Hardt et al. (2016))

## B Learning Implicit Cohorts

We learnt implicit cohorts by learning embeddings for individuals based on their historical utterances that are representative of their language usage and vocabulary usage. Each individual was considered as another token and appended at the beginning of the sentence which was fed to an LSTM based language model Sundermeyer et al. (2012) to train. These trained embeddings were then clustered using k-means clustering to extract cohorts.

## C Performance across cohorts

We look into the model performances across cohorts after adding the penalty term in the loss function. In order to check the consistency of our model, we created a spectrum of cohorts by randomly combining explicit and implicit cohorts which we had extracted before. For instance, we combined implicit cohort 1 (IC 1) and gender to create another cohort. The model performances over all the cohorts are shown in figures 1 2 3 . We observe the parity improve for almost all the cohorts as we increase $\lambda$, although we also note the decline in model performance for those cohorts which is a typical behavior after adding fairness constraints.

Figure 1: Model performances across cohorts in Yelp Dataset


Figure 2: Model performances across cohorts in Internal Dataset

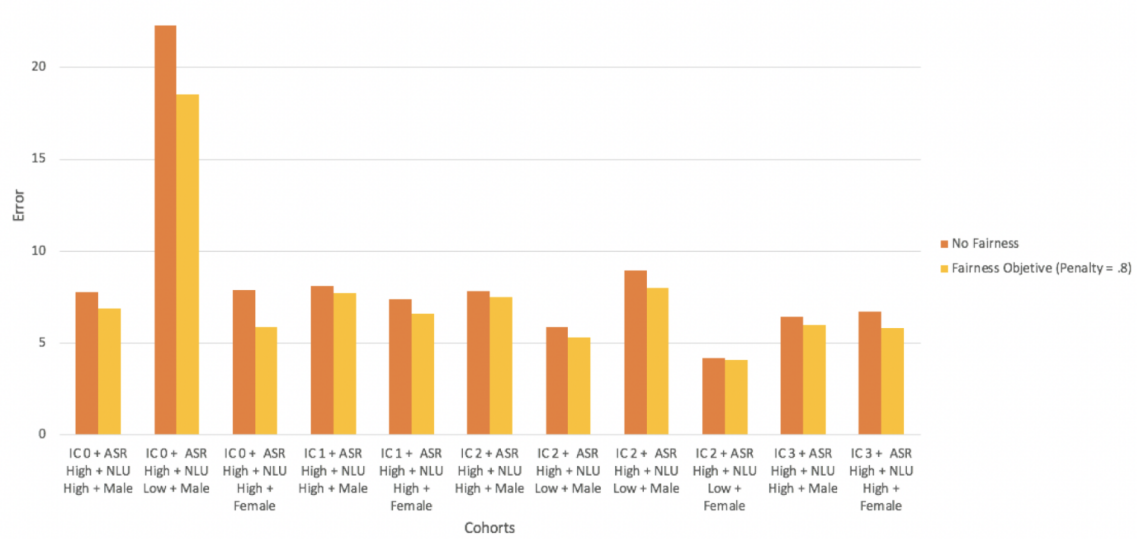


Figure 3: Model performances across cohorts in TrustPilot Dataset