AI/ML Assignment Report:
CS6603

Nehmya Negash
nnegash6@gatech.edu

1 AI/ML ASSIGNMENT REPORT

1.1 Subgroup to Protected Class Classification

Protected Class	Subgroup
Sexual Orientation	lesbian, gay, bisexual, transgender, trans, queer, lgbt, lgbtq, homosexual, straight, heterosexual
Sex	male, female, nonbinary
Race	black, white, african american, asian
Ethnicity	african, european, hispanic, latino, latina, latinx, mexican, canadian, american, indian, middle eastern, chinese, japanese
Religion	christian, muslim, jewish, buddhist, catholic, protestant, sikh, taoist
Age	old, older, young, younger, teenage, millennial, middle aged, elderly
Disability	blind, deaf, paralyzed

Table 1—Subgroup to Protected Class Classification

1.2 Calculating Toxicity Correlations

Protected Class	TOXICITY Correlation	Correlation Strength
Sexual Orientation	0.11	**Very weak**
Sex	0.017	**Very weak**
Race	0.06	**Very weak**
Ethnicity	0.15	**Very weak**

Religion	0.018	**Very weak**
Age	0.021	**Very weak**
Disability	0.012	**Very weak**

Table 2—Protected Class to Toxicity Correlation

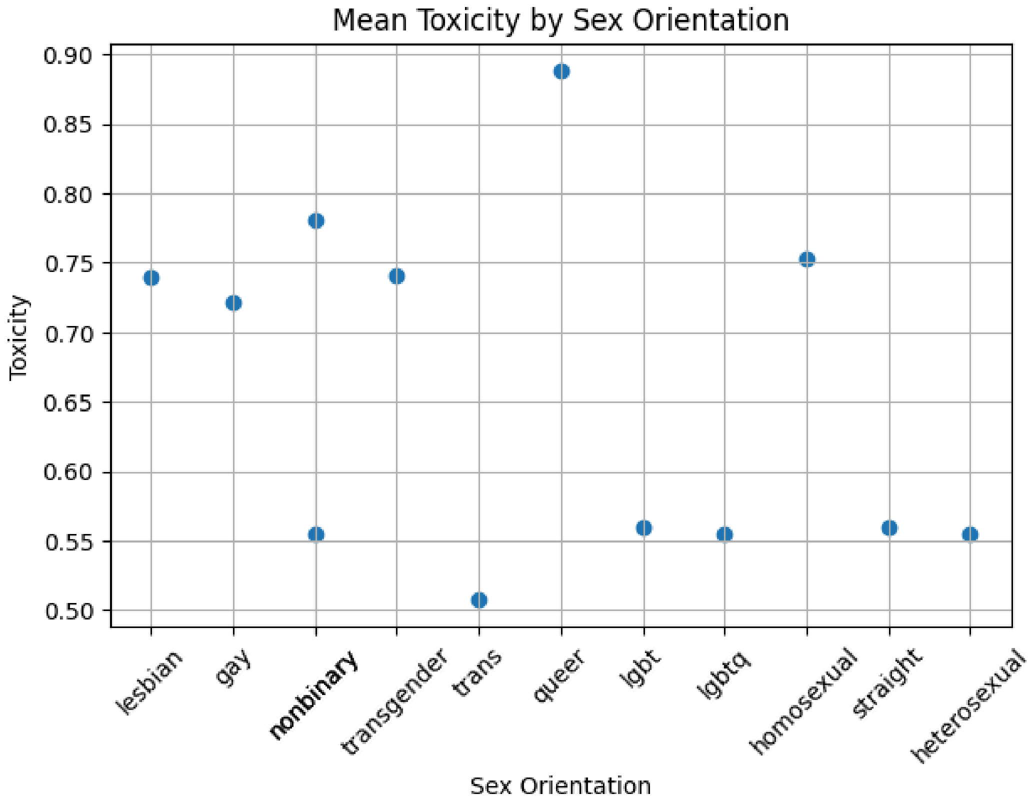

Figure 1—Sex Orientation vs Toxicity

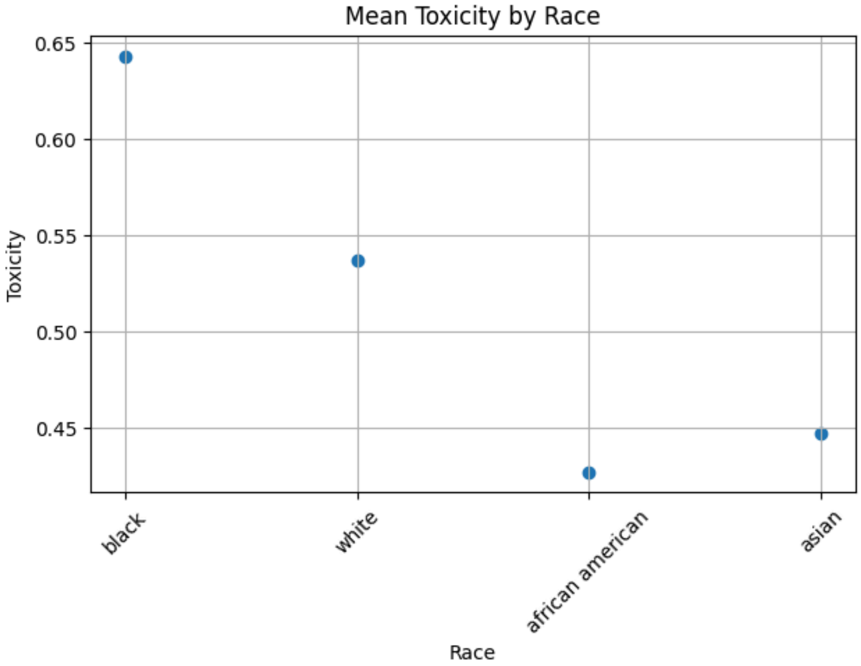

Figure 2—Race vs Toxicity

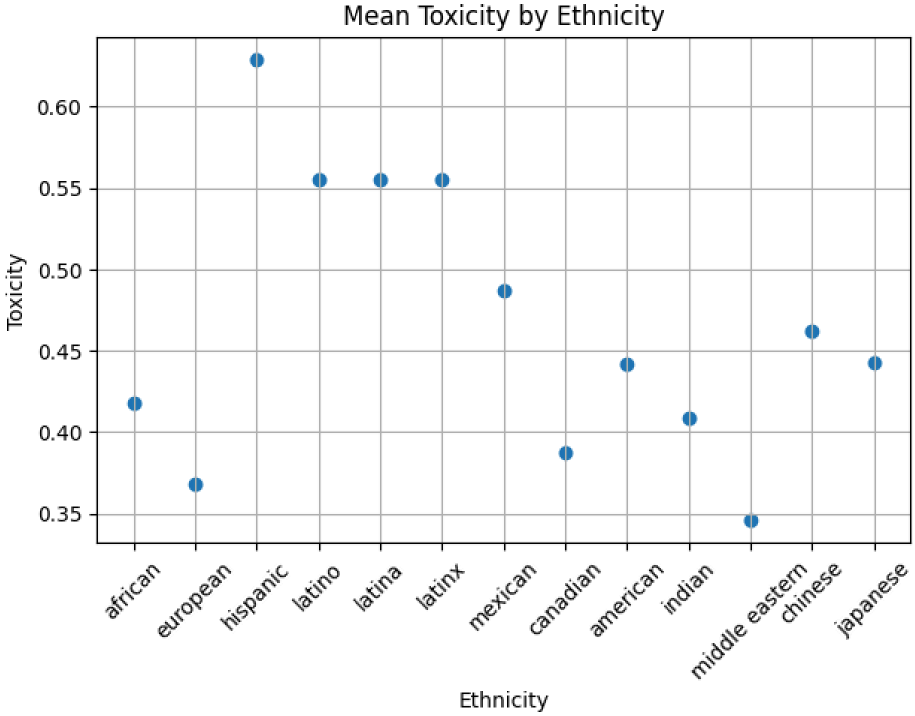

Figure 3—Ethnicity vs Toxicity

I agree that the correlations are quite weak. While some patterns do emerge across different groups, they aren't particularly meaningful. The differences in scores may seem noticeable at first glance, but they're actually quite small. That reflects my own experience as well. Outliers can sometimes give the impression that certain behaviors are tied to specific groups, but when you look more broadly across protected classes, those patterns don't really hold up.

1.3 Analyzing Toxicity on the Reduced Dataset

1.3.1 Total population

- mean: 0.5501
- std deviation: 0.3619
- range of 95%: [-0.1736, 1.2738]

1.3.2 Random 10% sampled

- mean: 0.5481
- std deviation: 0.3644
- MOE: 0.0081
- 95% Confidence Interval: [0.5399, 0.5562]

1.3.3 Random 60% sampled

- mean: 0.5493
- std deviation: 0.3623
- MOE: 0.0032
- 95% Confidence Interval: [0.5456, 0.5531]

1.4 Analyzing Toxicity for a Chosen Protected Class

I picked the Disability protected class that is made up of the "deaf", "blind", and "paralyzed" subgroups.

1.4.1 Total population

- mean: 0.5824
- std deviation: 0.3350

1.4.2 Random 10% sampled

- mean: 0.5861
- std deviation: 0.3343

- MOE: 0.0310
- 95% Confidence Interval: [0.5532, 0.6141]

Yes, the sample mean lies within the 95 confidence interval.

1.4.3 Random 60% sampled

- mean: 0.5852
- std deviation: 0.3356
- MOE: 0.0126
- 95% Confidence Interval: [0.5726, 0.5978]

Yes, the sample mean lies within the 95 confidence interval.

Both the 10% and 60% samples include the actual mean within their confidence intervals because they were calculated using a 95% confidence level. Since that's a fairly strong level of confidence, it's unlikely that the true mean would fall outside either interval. What really changes with the sample size is the width of the interval, not the confidence level itself. The 10% sample has a wider interval because there's more uncertainty with a smaller group—so the range expands to still capture the mean with 95% confidence.

1.5 Analyzing Toxicity for the Subgroups of the Chosen Protected Class

1.5.1 'blind' subgroup

1.5.1.1 Total population

- mean: 0.6369
- std deviation: 0.3080

1.5.1.2 Random 10% sampled

- mean: 0.6613
- std deviation: 0.3013
- MOE: 0.0449
- 95% Confidence Interval: [0.6028, 0.6994]

Yes, the sample mean lies within the 95 confidence interval.

1.5.1.3 Random 60% sampled

- mean: : 0.6242

- std deviation: 0.3098
- MOE: 0.0212
- 95% Confidence Interval: [0.6028, 0.6521]

Yes, the sample mean lies within the 95 confidence interval.

1.5.2 'deaf' subgroup

1.5.2.1 Total population

- mean: 0.5552
- std deviation: 0.3445

1.5.2.2 Random 10% sampled

- mean: 0.5824
- std deviation: 0.3289
- MOE: 0.0551
- 95% Confidence Interval: [0.5228, 0.6404]

Yes, the sample mean lies within the 95 confidence interval.

1.5.2.3 Random 60% sampled

- mean: 0.5554
- std deviation: 0.3434
- MOE: 0.0227
- 95% Confidence Interval: [0.5329, 0.5796]

Yes, the sample mean lies within the 95 confidence interval.

1.5.3 'paralyzed' subgroup

1.5.3.1 Total population

- mean: 0.5552
- std deviation: 0.3445

1.5.3.2 Random 10% sampled

- mean: 0.5524
- std deviation: 0.3412
- MOE: 0.0553

- 95% Confidence Interval: [0.4894, 0.6412]

Yes, the sample mean lies within the 95 confidence interval.

1.5.3.3 Random 60% sampled

- mean: 0.5553
- std deviation: 0.3429
- MOE: 0.0227
- 95% Confidence Interval: [0.5341, 0.5782]

Yes, the sample mean lies within the 95 confidence interval.

This is a fairly high confidence level, so there’s a strong likelihood that the true mean falls within the intervals for both the 10% and 60% samples. Since the sample size is factored into how the interval is calculated, each one is appropriately adjusted to maintain that level of confidence. As shown, smaller samples lead to wider intervals, reflecting the greater uncertainty that comes with less data.

1.6 Plots and Toxicity Analysis

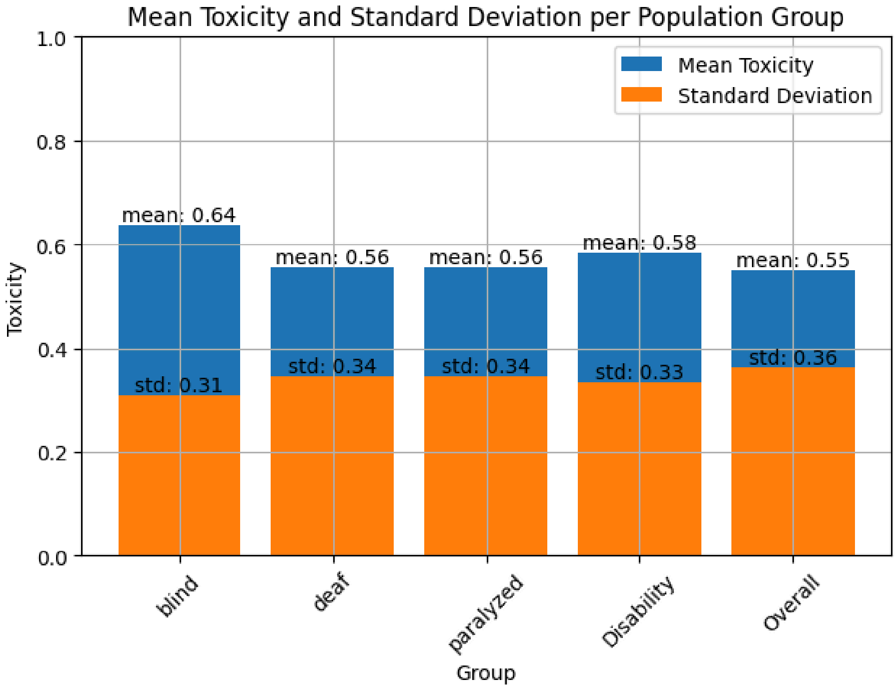

Figure 4—Population group Means and Standard Deviations

The ‘blind’ subgroup has the highest Toxicity mean value and the lowest standard deviation which means it consistently is around that mean value. The ‘deaf’ and ‘paralyzed’ subgroups have the same mean and standard deviation. These both have lower means and higher standard deviations compared to the ‘blind’ subgroup. The ‘blind’ subgroup has the largest difference in Toxicity value when compared to the population. It has a 0.09 greater mean and a 0.05 lower standard deviation. The combination of a higher mean with lower standard deviation provides a higher Toxicity value.

The human bias is shown in the data through how some subgroups are considered more toxic than others. This is shown through some of the slight correlations that we have seen with our analysis. Minimizing the bias can be dealt with by working on the algorithm that calculates the Toxicity value. Maybe an even distribution of subgroups across protected classes can create a more equally weighted set of groups where one is considered more toxic than the other.