AI/ML Part 1 Assignment Report

Juejing Han
jhan446@gatech.edu

1 AI/ML PART 1 ASSIGNMENT REPORT

1.1 Classify Subgroups into Protected Class Categories (Step 2)

In this section, columns in the dataset except for the first two columns ("Wiki_ID" and "TOXICITY") are classified into protected classes according to the following rules:

1) Columns falling under multiple protected classes, such as Chinese (considered as either race or national origin) or White (considered as either race or color), are assigned to one class. 2) Columns that could be classified into "Race", "National Origin", or "Color" are primarily placed under the protected class "Race." 3) Columns that could be classified as either "Gender Identity" or "Sexual Orientation," or fall under broader terms in these fields like LGBT, are primarily categorized under the protected class "Sexual Orientation."

Number of columns be categorized: 50

Number of Protected Class Categories: 6

Table 1 — Protected Class & Subgroups.

Protected Class (NO. of Members)	**Corresponding Members (Subgroups)**
Gender Identity (5)	male, female, nonbinary, transgender, trans
Sexual Orientation (9)	heterosexual, bisexual, homosexual, gay, lesbian, straight, queer, LGBT, LGBTQ
Race (17)	African American, Middle Eastern, Asian, European, Hispanic, Latino, Latina, African, Chinese, Japanese, Latinx, Mexican, Indian, American, Canadian, Black, White
Religion (8)	Christian, Catholic, Protestant, Muslim, Buddhist, Sikh, Jewish, Taoist
Age (8)	old, older, elderly, young, younger, teenage, middle-aged, millennial
Disability (3)	blind, deaf, paralyzed

1.2 Calculating Toxicity Correlations (Step 3)

Data cleaning: Remove 2 rows containing "nan" values in any column.

Data reducing: Remove 864 rows with "False" values across all subgroups.

Number of observations (rows) of original data: 76,565

Number of observations (rows) of reduced data: 75,699

For each protected class, determine the average toxicity of its subgroups and assign numerical values accordingly. For N numbers of subgroups, assign 1 to the subgroup with the lowest average toxicity, assign 2 to the subgroup with the second lowest average toxicity, and so on, up to N for the subgroup with the highest average toxicity.

Combine the numerical values of N subgroups (N columns) into one compacted column representing the entire protected class. If only one subgroup has a True observation (i.e., all other subgroups are False), the compacted column stores the assigned value of that subgroup. If multiple subgroups have True observations, the compacted column stores the average of the assigned values of those subgroups. The compacted column stores zero when all subgroups are False, indicating that the entire protected class has a False observation, which is excluded from the calculation of correlation coefficients between the protected class (the compacted column) and TOXICITY.

Classification Results – Protected Class Variables: As illustrated in Table 1.

Correlation Coefficients: As shown in Table 2. The correlation strength is assessed according to Evans's study (1996).

Table 2 — Correlation Coefficients between Protected Class and TOXICITY.

	Gender Identity	**Sexual Orientation**	**Race**	**Religion**	**Age**	**Disability**
TOXICITY	0.213	0.358	0.222	0.160	0.163	0.100
CORRELATION STRENGTH	Weak	Weak	Weak	Very weak	Very weak	Very weak

Sexual Orientation, Race, and Gender Identity exhibit the highest correlation coefficients with TOXICITY, all showing weak correlations. Fig. 1-3 show the numerical subgroup values (X-axis) versus average toxicity values for the three protected classes.

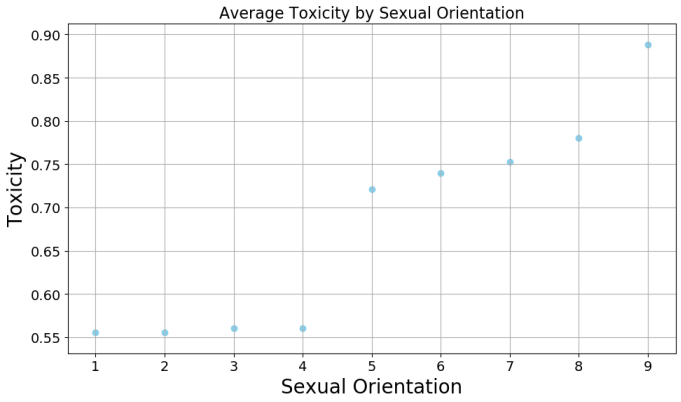

Figure 1 — Average Toxicity by Sexual Orientation

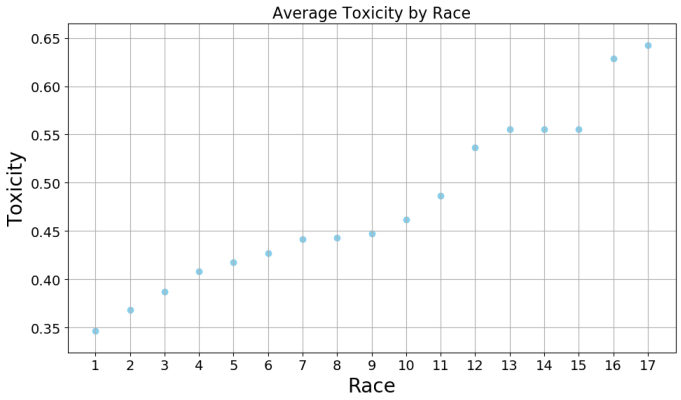

Figure 2 — Average Toxicity by Race

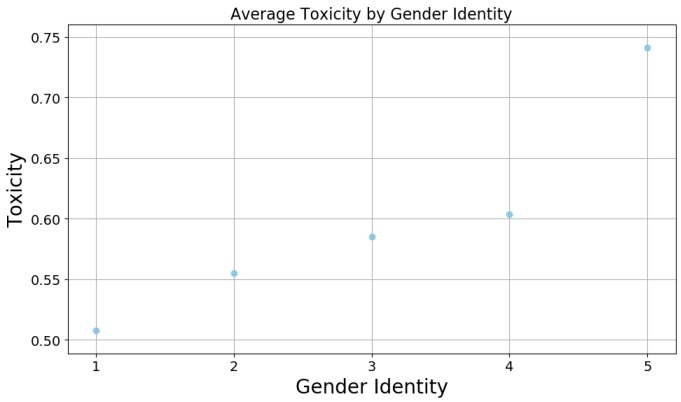

Figure 3 — Average Toxicity by Gender Identity

I agree with the correlation values listed in Table 2. The assigned subgroup values, sequentially ordered based on the average toxicity of each subgroup, exhibit a weak or very weak correlation with the toxicity values. This indicates little to no linear relationship between the assigned subgroup values and toxicity values. Consequently, a subgroup's high/low average toxicity value does not necessarily correspond to a high/low

toxicity outcome for an instance associated with that specific subgroup, suggesting an absence of significant bias based on the average toxicity values of subgroups.

1.3 Analyzing Toxicity on the Reduced Dataset (Step 4)

In this section, the range of toxicity values around the mean that covers 95% TOXICITY is calculated based on the formula:

$$Mean \pm 1.96 \times STDV$$

Where $Mean$ and $STDV$ are the mean and standard deviation of toxicity, respectively, and a Z-value of 1.96 corresponds to a 95% confidence interval[1].

The margin of error (MoE)[2] is determined as:

$$MoE = 1.96 \times STDV/\sqrt{n}$$

Where n denotes the sample size. The upper and lower confidence limits are expressed as $Mean \pm MoE$.

Randomly sample 10% and 60% of the reduced dataset.

1.3.1 *Toxicity in Reduced Dataset (Step 4.1)*

Mean: 0.550

Standard deviation: 0.362

Range of values around the mean that includes 95% TOXICITY: -0.159 to 1.259.

1.3.2 *Toxicity in 10% of Reduced Dataset (Step 4.2)*

Mean: 0.550

Standard deviation: 0.362

Margin of error: 0.0082

Lower confidence limit: 0.542

Upper confidence limit: 0.558

1.3.3 *Toxicity in 60% of Reduced Dataset (Step 4.3)*

Mean: 0.550

Standard deviation: 0.362

Margin of error: 0.0033

1 https://en.wikipedia.org/wiki/97.5th_percentile_point.

2 https://en.wikipedia.org/wiki/Margin_of_error.

Lower confidence limit: 0.546

Upper confidence limit: 0.553

1.4 Analyzing Toxicity for a Chosen Protected Class (Step 5)

Chosen protected class: Disability

Number of True observations (rows) of Disability: 4,542

In this section, the analysis conducted in section 1.3 is replicated for a selected protected class (Disability). The same statistics are calculated using the same datasets mentioned in section 1.3.

1.4.1 *Toxicity Associated with Disability in Reduced Dataset (Step 5.1)*

Mean: 0.582

Standard deviation: 0.335

1.4.2 *Toxicity Associated with Disability in 10% of Reduced Dataset (Step 5.2)*

Mean: 0.569

Standard deviation: 0.335

Margin of error: 0.0317

Lower confidence limit: 0.537

Upper confidence limit: 0.601

1.4.3 *Toxicity Associated with Disability in 60% of Reduced Dataset (Step 5.3)*

Mean: 0.581

Standard deviation: 0.337

Margin of error: 0.0126

Lower confidence limit: 0.568

Upper confidence limit: 0.593

1.4.4 *Sample Mean vs Population MoE (Step 5.4 - 5.5)*

a) For the 10% reduced dataset, the sample toxicity mean of the protected class Disability **does NOT lie within** the population margin of error.

b) For the 60% reduced dataset, the sample toxicity mean of the protected class Disability **does NOT lie within** the population margin of error.

The observed phenomena a) and b) could be attributed to **selection bias and inadequate sample size**.

1) Selection Bias: The selected sample (the chosen protected class) is not representative of the population. The distribution of toxicity within the protected class differs from the overall population distribution. The chosen protected class does not accurately represent the population, leading to discrepancies.

2) Inadequate Sample Size: The chosen protected class has a much smaller sample size compared to the overall population, comprising only approximately 6% of the total population. This smaller size may not adequately represent the population.

1.5 Analyzing Toxicity for the Subgroups of the Chosen Protected Class (Step 6)

In this section, the analysis conducted in section 1.3 is replicated for the subgroups of a selected protected class (Disability). The same statistics are calculated using the same datasets mentioned in section 1.3.

1.5.1 *Resulting Statistics (Step 6.1 - 6.3)*

Table 3 — Statistics for Subgroups in Disability.

Dataset	Subgroup	Mean	Standard Deviation	MoE	Lower Confidence Limit	Upper Confidence Limit
Reduced Data	blind	0.637	0.308	0.016	0.621	0.652
	deaf	0.555	0.345	0.017	0.538	0.573
	paralyzed	0.555	0.345	0.017	0.538	0.573
10% Reduced Data	blind	0.644	0.297	0.047	0.597	0.692
	deaf	0.530	0.358	0.059	0.471	0.589
	paralyzed	0.527	0.338	0.057	0.470	0.584
60% Reduced Data	blind	0.640	0.309	0.020	0.621	0.660
	deaf	0.540	0.349	0.023	0.517	0.563
	paralyzed	0.559	0.345	0.022	0.537	0.581

1.5.2 *Sample Mean vs Population MoE (Step 6.4 - 6.5)*

a) For the 10% reduced dataset, the sample toxicity mean of each protected class subgroup **does NOT lie within** the population margin of error.

b) For the 60% reduced dataset, the sample toxicity mean of each protected class subgroup **does NOT lie within** the population margin of error.

The potential reasons for a) and b) align with those mentioned in section 1.4.4.

1) Selection Bias: The distribution of toxicity for each subgroup (blind, deaf, or paralyzed) within the protected class differs from the overall population distribution. Each subgroup does not accurately represent the population, leading to discrepancies.

2) Insufficient Sample Size: Each subgroup has a considerably smaller sample size compared to the overall population, making up only about 2% of the entire population. This limited size may not provide a representative sample of the population.

1.6 Plots & Toxicity Analysis (Step 7)

Fig. 4 illustrates the mean and the standard deviation (STDV) of the population, the selected protected class "Disability," and each subgroup within the protected class. It indicates that among the three subgroups, **"blind" has the highest TOXICITY value**, while **"deaf" and "paralyzed" share the lowest value**, and **"blind" has the largest difference in TOXICITY value when compared to the population mean**.

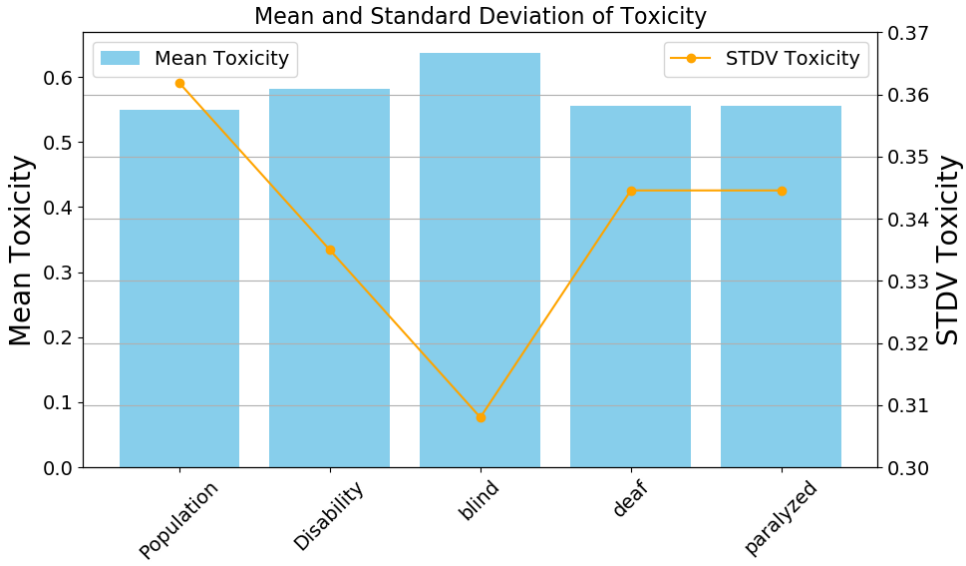

Figure 4— Mean and Standard Deviation of Toxicity

1.6.1 *Highest TOXICITY Value for "bind" (Step 7.1)*

Subgroup "blind" has the highest TOXICITY value, and the possible reasons are:

1) Biased Data: The assessment of TOXICITY could be influenced by bias stemming from human perceptions. Since the comments are labeled by human raters, subjective biases towards different disabilities can contribute to high TOXICITY values for the term "bind."

2) Prejudicial & Offensive Usage: The term "blind" could be used in contexts associated with prejudices against individuals with visual impairments. It could also be used in derogatory and offensive comments. These usages reflect societal biases and ableism, potentially resulting in high TOXICITY values for "blind."

1.6.2 *Lowest TOXICITY Value for "deaf" & "paralyzed" (Step 7.2)*

Subgroups "deaf" and "paralyzed" share the lowest TOXICITY value, and the possible reasons are:

1) Biased Data: Human perceptions, which might vary across different disabled conditions, may result in labeling low TOXICITY values for "deaf" and "paralyzed." This subjective bias can contribute to low TOXICITY values for these terms.

2) Reduced Offensive Usage: Unlike the term "blind," which often carries offensive connotations, "deaf" and "paralyzed" are typically more descriptive terms used in neutral contexts. This difference in usage could influence the labeling process, resulting in low TOXICITY values for "deaf" and "paralyzed."

1.6.3 *Subgroup "blind" has the Largest Difference in TOXICITY Value (Step 7.3)*

Subgroup "blind" exhibits the largest disparity in TOXICITY value when contrasted with the population mean. **The reasons for "blind" having the highest TOXICITY value align with those discussed in section 1.6.1.** Biased TOXICITY assessment may result in biased data, contributing to high TOXICITY values for "blind." Moreover, the term "blind" often carries societal biases and ableism, accompanied by offensive connotations, potentially leading to high TOXICITY values. These could also be the potential reasons why the subgroup "blind" exhibits the largest difference in TOXICITY value compared to the population mean.

1.6.4 *Human Bias (Step 7.4)*

According to the analyses in this report, there is a **human bias toward disability** in the data.

To quantify this bias, conducting a qualitative analysis of the dataset is essential. This involves examining patterns of bias and exploring demographic information. The statistics presented in this report serve as a good example, highlighting the disparities among different disabled conditions. They clearly indicate which subgroup in the Disability class has high or low TOXICITY values and the magnitude of the difference between them.

To minimize this bias, it is crucial to establish a neutral toxicity assessment framework that adheres to guidelines ensuring fairness and minimizing human bias in the rating process. Moreover, **ongoing monitoring and evaluation of the assessment process** can help identify and address emerging biases. Additionally, **comprehensive data selection procedures** should be implemented to eliminate biased samples. This includes careful consideration of the dataset composition and ensuring that all relevant demographics are adequately represented.

2 REFERENCES

1. Evans, J.D. (1996). Straightforward Statistics for the Behavioral Sciences. Thomson Brooks/Cole Publishing Co.