CS 6603: AI, Ethics, and Society
Homework Project #3: AI/ML Part 1

Readings:

- Dixon, Lucas & Li, John & Sorensen, Jeffrey & Thain, Nithum & Vasserman, Lucy. "Measuring and Mitigating Unintended Bias in Text Classification," AAAI/ACM Conference on AI, Ethics, and Society, pp. 67-73, 2018. https://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_9.pdf
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," NIPS 2016 -https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

In this assignment, you'll continue the process of exploring relationships in data. You'll accomplish this task by computing some basic inferential statistical measures on a natural language-based dataset.

Natural language processing is concerned with the ability to process and analyze large amounts of natural language data, whether for automated sentence completion in emails, conversational agents and chatbots, or AI tools to help journalists. In this assignment, we will work with data from a classifier built to identify toxicity in comments from Wikipedia Talk Pages. The model is built from a dataset of 127,820 Talk Page comments, each labeled by human raters as toxic or non-toxic. A toxic comment is defined as a "rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Step 1 – *Download the modified dataset available on CANVAS – toxity_per_attribute.csv*:

- <Wiki_ID> is unique identifier associated with Wikipedia comment
- <TOXICITY> is a toxicity value from 1 if the comment was considered toxic and value 0 if the comment was considered neutral or healthy
- < subgroup > columns: One column per human attribute; True if the comment mentioned this identity.
- Due to sensitivity, comments were removed to construct the modified dataset. The original data source can be found at: https://github.com/conversationai/unintended-ml-bias-analysis/tree/master/data

Step 2: Classify Subgroups into Protected Class Categories

1. Identify the protected class each subgroup belongs in (e.g. Christian, Muslim, X belong to the protected class Religion).
2. If a subgroup falls under two protected classes, you may keep it in both protected classes or classify it under only one of the protected classes (e.g., Chinese could fall under Race or National Original. You can choose both or only one).
3. In your report, provide a list of the protected class categories and their corresponding members.

Step 3: Calculating Toxicity Correlations

1. Create a *reduced data set* by deleting any rows that have all FALSE values for every column in that row.

   Note: This is the *reduced data set* that will be used in all subsequent steps.

2. Using the reduced data set, identify an objective ordering scheme for each protected class category by defining values for each of its protected class members.

   To do this, convert FALSE values to 0 and TRUE values to a unique numerical value for each

subgroup member (e.g., for gender identify: FALSE = 0; male = 1; female = 2; binary = 3; etc.).

You may also combine group members and assign numerical values based on your belief about similarities among the group members (e.g., gender identify: FALSE = 0; all others = 1; female = 2).

3. Using your assigned numerical values, create a compacted data set by combining the subgroup into one column representing the protected class category (e.g. combine all columns related to Religion into one Religion column).

4. Calculate the correlation between the protected class columns and TOXICITY.

   In a table, provide the correlation coefficients for each protected class and the strength of the correlation.

   Note: As guidance, you can use (Evans, J. D. (1996). Straightforward statistics for the behavioral sciences. Brooks/Cole Publishing) which suggests the following related to the absolute value of the correlation coefficient:
   - .00-.19 “very weak” correlation
   - .20-.39 “weak” correlation
   - .40-.59 “moderate” correlation
   - .60-.79 “strong” correlation
   - .80-1.0 “very strong” correlation

5. Using the mappings from Step 3.2, plot the numerical Subgroup values vs Toxicity for the three highest correlation coefficients.

   In your report, answer the following: *Do you agree with the correlation values? Why or why not?*

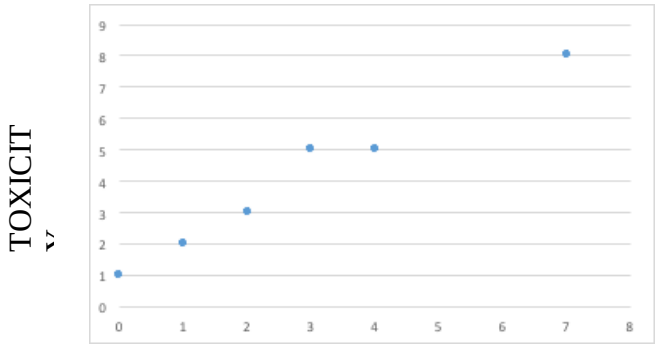
Step 3 Example Output (for illustrative purposes only):

Classification Results - Protected Class Variables:
- Religion: Christian, Muslim
- Age: younger, older

Correlation Coefficients:

	Religion (Protected Class Variable)	Age (Protected Class Variable)
TOXICITY	0.03	0.7
CORRELATION STRENGTH	Very weak correlation	Strong correlation

TOXICITY and Age are strongly correlated.



Step 4: Analyzing Toxicity on the Reduced Dataset

1. Using your reduced dataset (from the first task in Step 3), calculate and report the population mean and population standard deviation of TOXICITY. List the range of values around the mean that includes 95% TOXICITY.
2. Run the random sampling method using 10% of the data. Report the mean, standard deviation, and margin of error.
3. Run the random sampling method using 60% of the data. Report the mean, standard deviation, and margin of error.

IMPORTANT: *For the next steps below, choose a protected class category for analysis.*

Step 5: Analyzing Toxicity for a Chosen Protected Class

1. Using the reduced dataset (Step 3.1), calculate the mean and standard deviation of TOXICITY associated with the protected class.
   Hint: TOXICITY values should only be included in the calculation when the associated protected class value is not FALSE.
2. Using the 10% randomly sampled dataset (Step 4.2), calculate and report the mean, standard deviation, and margin of error for the protected class.
3. Using the 60% randomly sampled dataset (Step 4.3), calculate and report the mean, standard deviation, and margin of error for the protected class.
4. Using the 10% reduced dataset, indicate (Yes/No) if the sample mean of the protected class (Step 5.2) lies within the population margin of error (Step 4.2). Repeat for the 60% reduced dataset.
5. Explain why you think the 10% dataset protected class sample mean does or does not lie within the 10% dataset population margin of error. Repeat for the 60% dataset.

Step 6: Analyzing Toxicity for the Subgroups of the Chosen Protected Class

1. Using the reduced data set (Step 3.1), calculate the mean and standard deviation of TOXICITY associated with each subgroup that is a member of the protected class selected in Step 5.
   Hint: TOXICITY values should only be included in the calculation when the associated protected class value is not FALSE.
2. Using the 10% randomly sampled dataset (Step 4.2), calculate and report the mean, standard deviation, and margin of error for each subgroup of the protected class.
3. Using the 60% randomly sampled dataset (Step 4.3), calculate and report the mean, standard deviation, and margin of error for each subgroup of the protected class.

4. Using the 10% reduced dataset, indicate (Yes/No) if the sample mean of each protected class subgroup (Step 6.2) lies within the population margin of error (Step 4.2). Repeat for the 60% reduced dataset.
5. Explain why you think the 10% dataset protected class subgroup sample means do or do not lie within the 10% dataset population margin of error. Repeat for the 60% dataset.

Step 7: Plots and Toxicity Analysis

Plot on one graph the following:

1. the computed population mean and standard deviation (Step 4)
2. the computed mean and standard deviation for the protected class category (Step 5)
3. the computed mean and standard deviation for each subgroup of the protected class category (Step 6).

Answer the following questions. Justify your reasoning.

1. Which subgroup has the highest TOXICITY value? Explain your reasoning.
2. Which subgroup has the lowest TOXICITY value? Explain your reasoning.
3. Which subgroup has the largest difference in TOXICITY value when compared to the population mean? Explain why you think this group has the HIGHEST toxicity value.
4. What type of human bias is in the data?
    - How can you quantify this bias?
    - How can it be minimized?

Step 8: Submission

Turn in a report (in PDF format) documenting your outputs for each question. The report should follow the JDF format. You can find a link to the JDF template here: JDF Templates.

We suggest using the Microsoft Word template for proper formatting and styling. Reports that are not neat and well organized will receive up to a 10% deduction.

The file name for submission is GTuserName_Assignment_3, (ex. gBurdell3_Assignment_3.pdf). Deductions will be made if your file name is not submitted correctly.