AI, Ethics, and Society

Spring 2020

Homework Project #3

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. http://www.aies-conference.com/wp-content/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 – 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: - Identify the protected class categories and assign numerical values for each member associated with each protected class category by:

- Combining values from the related subgroups into the relevant protected class category (e.g. christian + muslim + X -> Religion)
- To define values for each protected class member, convert FALSE to 0 and TRUE to a unique numerical value for each subgroup member (e.g. for gender identify: FALSE = 0; female = 1; binary = 2; etc.)
- Provide the classification results (i.e. list of protected class categories and their associated protected class members)

Step 3:

- Create a reduced data set by blending columns related to the protected class categories and deleting any rows that have all FALSE values for every column in that row
- Calculate the correlation between the protected class category and TOXICITY. Provide the correlation coefficients in table format and identify the strength of the correlation. Select the three highest correlation coefficients and plot data for the correlated variables; indicate its correlation strength [Note: there may/may not be any strong correlations in this dataset].
- As guidance, 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

Example Output:

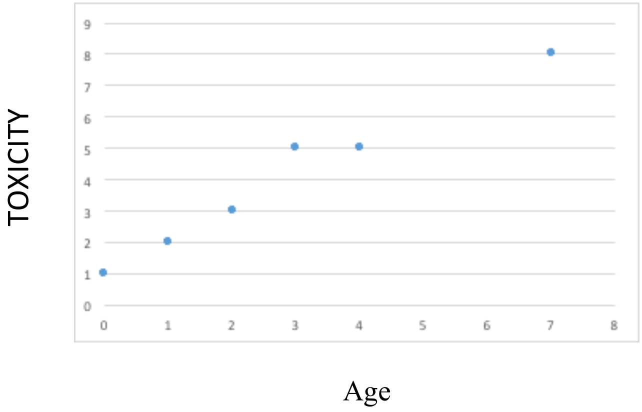
Classification Results - Protected Class Variables:

- 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: Calculate the population mean and population standard deviation of TOXICITY. What is the range of values around the mean that includes 95% of TOXICITY? Run the random sampling method using 25% and 50% of the data. For each, what is the mean, standard deviation, and margin of error?

For these questions, choose only one of the protected class categories.

Step 5: Calculate the mean and standard deviation of TOXICITY associated with the protected class category (***Hint:*** TOXICITY values should only be included in the calculation when the associated protected class value is not FALSE). Run the random sampling method using 25% and 50% of the data. For each, what is the mean and standard deviation? Indicate (yes/no) if the values lie within the associated population margin of error.

Step 6: Calculate the mean and standard deviation of TOXICITY associated with each subgroup that is a member of the protected class category (***Hint:*** TOXICITY values should only be included in the calculation when the associated protected class value is not FALSE). Run the random sampling method using 25% and 50% of the data. For each subgroup, what is the mean and standard deviation? Indicate (yes/no) if the values lie within the associated population margin of error.

Step 7: Plot (on one graph) - 1) the computed population mean/standard deviation (Step 4), (2) the computed mean/standard deviation for the protected class category (Step 5), and (3) the computed mean/standard deviation for each subgroup of the protected class category (Step 6). Which of the subgroups has the highest TOXICITY value? Which of the subgroups has the lowest TOXICITY value? Which of the subgroups has the largest difference in TOXICITY value when compared to the population mean? Does there seem to be any human bias in the data? Explain (in no more than 3-5 sentences).

Step 8: Turn in a report documenting your outputs.