WHEN THE GROUND TRUTH IS NOT TRUE

A Study of Perturbations on “True” Toxicity in Text Classification

Neha Gupta, Yujing Ke, Neil Pruthi, Brandon Fain, Ashwin Machanavajjhala

Motivation

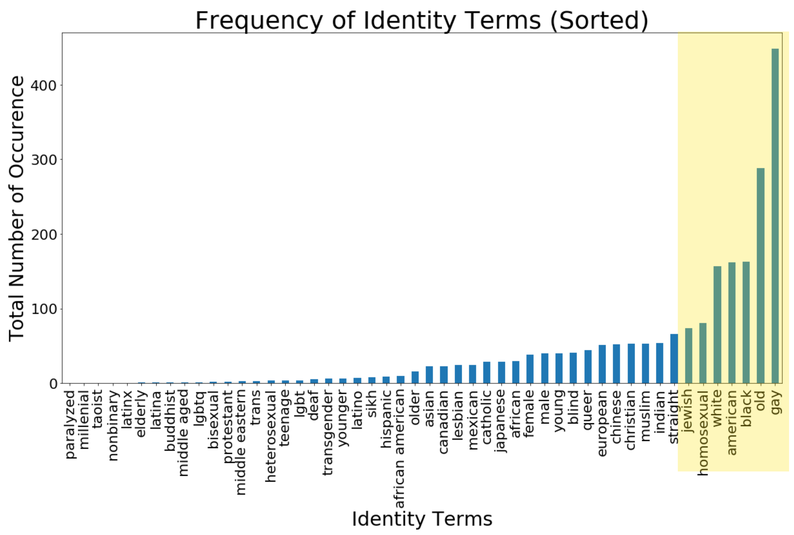
Posting comments online is easy, but ensuring comments are appropriate is difficult. Negative online behavior can drive away and upset users. Websites may want to deploy a classifier to automatically flag comments that are mean, hateful, prejudiced, offense, or some combination thereof.

However, a classifier may exhibit bias and lead to unfairness because it may predict, correctly or incorrectly, a toxic label for comments about a particular demographic group more often than comments about other demographic groups. This can stem from biased input data, because the definition of a toxic comment is a subjective opinion. We are interested in how different levels of dataset bias affect the bias of a classifier trained on that dataset for text classification, a context where crowdsourced data makes it difficult to discern ground truth.

Method

We use a dataset of around 100,000 Wikipedia Talk Page comments that have been labelled by humans as *toxic* or *non-toxic*. We used a logistic regression classifier to predict whether a comment is *toxic* or *non-toxic* on the test dataset.

In order to ensure we have sufficient sample sizes for each identity term, we only show results for demographic identity terms that appear in at least 3% of comments containing identity terms. The figure below shows the identity terms, in set T, that met this criterion.



Method (continued)

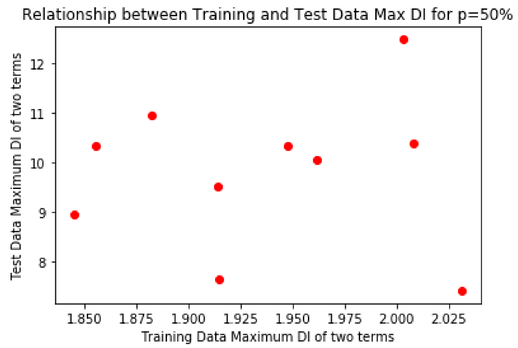
Random Variations in Toxicity of Training Data

To introduce noise into the training data, we randomly select comments with some probability, *p*, and for those comments, we assign a random toxicity value of either [0,1], where 1 indicates that a comment is *toxic*.

Our measure for bias in data is disparate impact, generalized from its typical two-class definition. Intuitively, a high disparate impact value means comments containing term t_1 are more likely to be labeled as toxic than comments containing a term t_2. We compute disparate impact for all pairs of terms and find the pairwise max pairwise:

$$DI_{T:[t_1,t_2,...t_n]} = max\left(\frac{P(comment\ containing\ t_i\ is\ toxic)}{P(comment\ containing\ t_j\ is\ toxic)}\right)$$

The figure below shows disparate impact of the training data, and predicted toxicity of the test data.

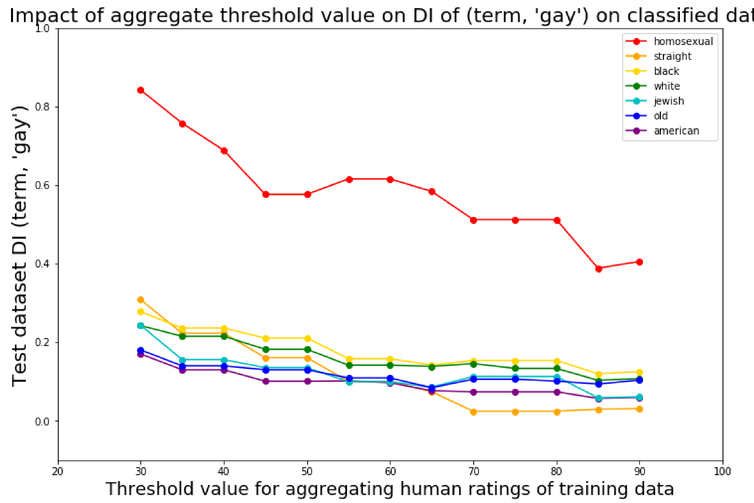


Variation in Cautiousness of Toxicity Definition

The original dataset of crowdsourced comment toxicity ratings has around ten human raters for each comment [2]. [1] labels a comment as toxic in the training dataset when the mean of the toxicity ratings across human raters is greater than 0.5. We adjust this threshold and observe changes in the disparate impact. Intuitively, a low threshold errs on the side of caution, and labels something as toxic if even few of the human raters rate it as toxic. Thus, a low threshold has more comments labelled as toxic.

We compare each demographic identity term to the term ‘gay’, as shown in the following figure. We choose to use the term ‘gay’ as it occurs most frequently, and is used as a metric in [1].

For the majority of terms, as threshold increased, DI ([term],’gay’) decreased for both the training data and the test datasets. In the next figure, we display the test dataset’s DI for each term at each threshold, to show how much change there is between bias of the two terms after training a classifier.



Results

Our work illustrates how text classifiers introduce biases between demographic groups, even when training data has randomly minimized bias, and that this bias can be either exacerbated or minimized when varying human rater consensus.

A. The graph showing the relationship between training and test DI shows that regardless of the percentage of noise introduced to the dataset, the disparate impact of the test data is unaffected by a range of training data disparate impacts. The bias of predicted values seems unaffected by random noise introduced in training data.

B. The results of the threshold graph shows that when we are less cautious in defining a consensus for toxicity, by requiring a higher threshold of mean toxicity, the probability that terms are predicted to be similarly toxic to ‘gay’ tends to decrease.

Further Work

- We hope to see whether additional fairness metrics such as Equality of Opportunity and Predictive Parity have a relationship between varied training data
- We hope to look into a relationship between selection of demographics of the human raters and the resulting disparate impact of various identity pairings

References

1. Dixon, Lucas, et al. "Measuring and mitigating unintended bias in text classification." available at: www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_9. pdf 2018.
2. Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. "Ex machina: Personal attacks seen at scale." Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017.

Department of Computer Science
Duke University
cs.duke.edu