Jennifer Wright
Senior Design Writing 1 Revised
Due: 10/1/2019

Proposal for Bias Detection Web Plugin

Today, we are lucky to live in a world where people care about moving towards equality. To help guide this push there are unlimited resources on the world wide web. While this intention is admirable, studies have shown that human bias often times slips into the text corpora from the world wide web. According to Caliskan in her report *Semantics derived automatically from language corpora contain human-like biases*, "text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names." While bias against insects may not seem harmful, bias towards race and gender is problematic and indicates an implicitly biased set of resources on the world wide web. Eliminating bias on the internet will be challenging and nearly impossible when it is not clear where it exists. The bias detection web plugin aims to help with this. With our product, we aim to bring awareness to the bias that surrounds us on the web every day in a visual and easily interpretable way. Our tool will use machine learning to extend the research that Professor Caliskan has done, using the Implicit Association Test to replicate biases and detect them. The tool will act as a litmus test, highlighting potentially biased phrasing and giving a general summary of the amount of bias in the work. Our hope is that with more awareness of the bias in work we rely on, content consumers will be able to recognize problematic biases and content creators will be held accountable to create less biased work in the future.

As with any project, there will be a few technical challenges that we foresee. Although we would like to think otherwise all human beings have some implicit bias based on their environment, and apart from obviously biased statements, it is difficult to determine a standard test to determine what is and what is not biased. This is an obstacle unique to our project because we are trying to define a model that can predict the bias when we as humans are not great at detecting it in the first place. That is why our model will be based off of the Word-Embedding Association Test, a statistical test developed by Aylin Caliskan and her team during her research on the subject based off of the Implicit Association Test. While this test may also have minor flaws, it is overall generally well received and respected. Next, finding training data that is reliable and robust enough to train a model is always challenging. We know that the corpora from the web is biased so we will use this to train our model while being careful that we are correctly making the identification. Performing extensive testing will also give us assurance that our algorithm is working the way we expected. Finally, overfitting is always a risk with machine learning, we want to make sure that we do not over-detect bias in text giving people a false positive of the amount of bias on the web. We will develop our web plugin consciously knowing that overfitting is a possibility. Having large amounts of descriptive data will also help our model be more

accurate. Exhaustive testing will also ensure accuracy -- we have the entire web to use as test data for this algorithm.

Now that the impetus for our product has been defined and some technical challenges addressed, I would like to present a potential use case. One of the reasons our team wants to develop this product is because it has been found that males are more associated with the subjects of math and science while females are more associated with performing well in verbal subjects. An interesting study performed at Stanford asks the question “why focus on the gender gap in math where males outperform and outnumber females over the gender gap in other subjects where females outperform and outnumber males?” this is because math, unlike other subjects has been shown to be a good predictor of future income according to *Explaining the Gender Gap in Math Test Scores: The Role of Competition* by Niederle and Vesterlund. Young women are learning from our biased web, seeing males being associated with math and thinking there is no one like them in math and thus it is not a place for them. Before young women know if they enjoy math, she could be turning away from it. This could indicate a potential butterfly effect of bias leading to a pay gap. Being aware of the issue and how broad it is can be the first step in fixing it. Our web plugin for bias detection accomplishes this goal.

Resources

Semantics derived automatically from language corpora contain human-like biases by Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan

Explaining the Gender Gap in Math Test Scores: The Role of Competition by Muriel Niederle and Lise Vesterlund