Spring 2019

Lab Practicum For Bias In Algorithms

Ameet Soni
Swarthmore College, soni@cs.swarthmore.edu

Krista Karbowski Thomason
Swarthmore College, kthomas2@swarthmore.edu

Lab Practicum

Due: Monday, April 8, 2019 in class

Instructions: Turn in your written response individually. For Part 1, you should design and conduct your experiment as a pair but write your responses independently. You may sure the results (i.e., the numerical findings) with each other. Please indicate who you worked with in your response. Part 2 should be done independently.

Part 1: Bias in word embeddings (1-1.5 pages)

Using the tools provided for testing associations with pairs of words, explore the implicit bias that can be encoded in natural language word embeddings. First, begin by replicating some of the pairings from the readings (e.g., European names vs. African names when paired with pleasant/unpleasant words). Then, respond to the following two prompts.

1) Rerun the pairings by varying the underlying training corpus used to learn the word embeddings. Discuss what impact, if any, this has on introducing biases into the trained word embedding model. Be sure to try each of the three datasets on both a task related to race/ethnicity and a task related to gender. Is there a noticeable change in each case?
2) Propose a new set of WEAT pairings to see if there are additional biases (e.g., religion, nationality, class). Generate a list of words for your target pair and use one of the existing attribute pairs (unpleasant/pleasant, family/career, male/female) or create your own. Describe your hypothesis (including how you created your lists of words) and then your findings. Discuss the implications of your results (e.g., what does this tell us about the algorithm, training corpus, and/or the experimental design itself).

Part 2: Ethical case study (2 pages)

Consider the following scenario you have been asked to review by a regional hospital:

> The emergency room has had difficulties with their triage system for prioritizing patients - it is both labor intensive (nurses spend less time caring for the patient and more time filling out paperwork and assessing the severity of a patient's case) and error prone (incorrect prioritization can cause prolonged suffering and worsening health). The hospital would like to use natural language processing to suggest prioritization for a patient. The patient submits a description of their condition which can be supplemented by a medical assessment (e.g., paramedic or nurse) and the system outputs a triage score. The system undergoes significant training using notes/scores from triage nurses on prior patients.

Identify *ethical* issues you believe need to be addressed by the hospital and/or developers. You can use relevant readings from class to help you craft your response. It's better to choose one

or two of the most important concerns and explain them in detail rather than try to cover too many things in this short assignment. For each concern, be sure to a) state your concern, b) justify the concern (i.e., why the issue is a possibility, including referencing our readings) and c) explain the potential ethical implications if the issue is ignored.