Repository Entry
Embedded EthiCS @ Harvard Teaching Lab

Overview

Course:	AC 295: Deep Learning for NLP
Course Level:	Graduate
Course Description:	"How can computers understand and leverage text data and human language? Natural language processing (NLP) addresses this question, and in this course students study the current, best approaches to it. No prior NLP experience is needed, but it is welcomed. This course provides students with a foundation of advanced concepts and requires students to conduct significant research on an NLP topic of their choosing. The aim is to produce a short paper worthy of submitting to an NLP conference. Assessment also includes pop quizzes, homework assignments, and an exam. The course starts with language representations and modelling, followed by machine translation (converting text from one language to another). Next, students learn about transformers (e.g., BERT and GPT-2), which are incredibly powerful deep learning models that currently yield state-of-the-art results in nearly every NLP task. We end the semester by covering tasks concerning bias and fairness, adversarial approaches, coreference resolution, and commonsense reasoning."
Module Topic:	Embedding Bias
Module Author:	Ellie Lasater-Guttmann
Semesters Taught:	Fall 2021
Tags:	word embeddings [CS], training data [CS], NLP [CS], bias [phil], identity [phil], representational harm [phil], allocative harm [phil]
Module Overview:	This module reviews instances of bias in AI, NLP, and then word-embeddings with the goal of identifying when design choices can cause representational and/or allocative harms. These harms ground the decision that these models should be debiased. Then we conclude by discussing what debiasing would look like.
Connection to Course Material:	Up until the point this module ran, the course covered the technical components of NLP, including word-embeddings. This module covers gender bias in word-embedding models and then reviews their application to see how that bias leads to harm. The course then moves to discuss applications of NLP and word-embeddings, so this acted as a bridge to that new material.
	We chose this topic based on the vast technical material about this type of bias already available. Future versions of this module could include a focus on debiasing for at least half of the course time.

Goals		
Module Goals:	- Identify instances of gender bias in the world, AI, NLP, and then word-embedding models - Understand the difference between representation and allocative harms - Identify when instances of gender bias lead to these two harms - Judge whether debiasing is worthwhile - Brainstorm how to debias these models	
Key Philosophical Questions:	1. When is bias harmful? 2. When (and why) are biased results harmful in word-embedding software? 3. What type of harm is caused by gender bias in NLP? 4. Are debiasing efforts worthwhile? 5. How would we judge the efficacy of debiasing efforts?	In the future, I would reconsider question 1 and instead focus time on questions 4 and 5. Question 1 ended up being too broad to remain engaging throughout the module time.

Materials		
Key Philosophical Concepts:	● Bias ● Representational Harm ● Allocative Harm	Working through the concept of bias (broadly in the world and then more narrowly in NLP) allows students to see that bias leads to representational and allocative harms. These harms then motivate debiasing.
Assigned Readings:	● Bolukbasi, T. et al. (2016) "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings" ● Gonen, Hila. Goldberg, Y. (2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them"	These two papers provided both the technical material and the ethical background required to understand the problem at issue. Bolukbasi et al. uncovers statistical gender bias in word embeddings. Gonen & Goldberg argues that efforts to debias embeddings have been insufficient. They are also the two most commonly cited papers on the topic.

Implementation

Class Agenda:	1. Review instances of gender bias in the world and AI 2. Understand that these instances would cause representational harms 3. Review other instances of gender bias in the world and AI 4. Understand that these instances would cause allocative harms 5. Review gender bias in NLP and identify harms 6. Review gender bias in word-embeddings and discuss harms in small groups 7. Evaluate whether debiasing is worthwhile 8. Brainstorm what debiasing would look like (in small groups and then as a class)	The larger discussion about gender bias in the world and AI broadly could be shortened to accommodate a more diligent discussion about debiasing. Students understood that bias was rampant and were interested in solving the problem even before we moved to debiasing.
Sample Class Activity:	Two small group discussions: 1. What applications of word-embeddings could lead to allocative harms? What applications of word-embeddings could lead to representational harms? 2. What would a debiased word-embedding model look like? What steps would you take to debias? Would you alter the corpus or debias after training?	Students were independently interested in whether debiasing is the responsibility of the software designer or instead the responsibility of those in the world / those who contribute to the corpus. I would recommend engaging this question directly.
Module Assignment:	Given that this is a graduate course that regularly participates in open-ended essays, they were assigned the following essay topic: You have designed a word-embedding model based on a real-life corpus. You feel confident that the model is debiased (at least in terms of gender). (1) How could you be sure that your model is debiased? (2) Could your model still cause allocative and/or representational harm?	These questions build well from the module in which we motivated debiasing efforts. Now, on their own, students need to evaluate what said efforts would look like.
Lessons Learned:	1. Revamping the activity to be more interactive would be useful (possibly with a choose-your-own-adventure). These students will now put together research projects and an activity that mirrored research projects could be a good match to the course content.	

2. Less time could be taken reviewing gender bias in the world/AI broadly and more time could be spent on debiasing.