Repository Entry
CS 287: Natural Language Processing

Course Level:
Graduate

Course Description:
"Machine learning for natural language processing with a focus on deep learning and generative models. Topics include language modeling, information extraction, multi-model applications, text generation, machine translation, and deep generative models. Course is taught as a reading seminar with student presentations." (Course Description)

Module Topic:
Bias and Stereotypes in Word-Embedding software

Module Author:
Diana Acosta-Navas

Semesters Taught:
Spring 2019

Tags:
- natural language processing (CS)
- word embeddings (CS)
- machine learning (CS)
- bias (phil)
- stereotypes (phil)
- discrimination (phil)

Module Overview:
The module examines the relation between gender stereotypes and the biases encoded in word embeddings. Students discuss some of the ethical problems that arise when gender bias becomes encoded in word embeddings, including the perpetuation and amplification of stereotypes, the infliction of representational and allocative harm, and the solidification of prejudice. After discussing some pros and cons of debiasing algorithms, the final part of the module explores the moral concerns that this solution may raise. This final discussion

focuses on the thought that bias often happens without our full awareness, hence debiasing and other technical solutions should be immersed in wide-ranging cultural transformations towards inclusion and equality.

Connection to Course Technical Material:

In the lead-up to the module, the course covers word-embedding techniques and their potential uses in processing natural language. In the module we examine a potential drawback of these techniques and the ethical problems raised by their employment, while also examining the advantages and disadvantages of alternative approaches. Specifically, the module invites students to weigh the technical advantages of word-embeddings against their potential to propagate gender stereotypes by encoding biases rooted in our use of language. Students are provided with philosophical concepts that help them articulate whether taking advantage of the computing power offered by word embeddings justifies the kind of harm that may be inflicted when biases are perpetuated and solidified.

Module Goals:

1. Introducing students to the concepts of bias, stereotypes, and discrimination.
2. Discussing the existence of gender biases in word-embedding software, and its correlation to gender stereotypes.
3. Guiding students in thinking about the ethical problems raised by the presence of gender bias in word-embedding software.
4. Prompting students to consider the potential advantages of debiasing word-embeddings, and its potential drawbacks.
5. Using case-studies to train students to identify morally problematic aspects in the context of complex real-world scenarios.

Key Philosophical Questions:

1. What are the distinctive features of stereotypes?
2. What makes stereotypes morally problematic?
3. Can individuals be harmed by the presence of stereotypes in language processing software?
4. Can debiasing algorithms resolve the issue given that bias and stereotypes are widespread in our culture?

Key Philosophical Concepts:

- Bias (explicit vs. implicit)
- Stereotypes
- Discrimination
- Statistical truths vs essentialist claims
- Prejudice
- Representational vs. allocative harms

Assigned Readings:

Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V. and Kalai, A.T., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357). https://arxiv.org/abs/1607.06520

⇒ This piece shows that word embedding software trained on Google News articles exhibits female/male gender stereotypes and argues that the widespread use of this software could potentially amplify whatever biases are coded in their data. It suggests a methodology for modifying embeddings in a manner that removes gender stereotypes without sacrificing the computational power of these algorithms.

Class Agenda:

1. Active learning exercise: identifying analogies that reflect gender stereotypes.
2. Class discussion: what is a stereotype?
3. Presentation on the findings of gender biases in word2vec.
4. Small group discussion: what is wrong with allowing gender biases into word-embeddings?
5. Class-wide discussion about the moral issues raised by different kinds of bias.
6. Discussion of debiasing techniques and their advantages and disadvantages.

Sample Class Activity:

At the beginning of the session, students are given a list of analogies that link professions to genders, including ballerina/dancer, hostess/bartender, vocalist/guitarist, among others. They are asked to mark those analogies that reflect gender stereotypes. When they finish,

the lecturer polls students to find out how they responded to four analogies: one that is clearly stereotypical (homemaker/computer scientist), one that is not (Queen/King), and two that are debatable (Diva/Rockstar, and Interior Designer/Architect). The Embedded Ethics fellow then leads a discussion about the distinctive features of gender stereotypes, which serves as a starting point to discuss the ethical problems raised by the existence gender biases in word-embeddings.

Assignment:

Students are asked to imagine that they are tasked with producing an image captioning software that employs machine learning. They are directed to focus on the generation of gender-specific caption words, choosing between two models. The first model relies on learned priors based on the image context. It exploits contextual cues to determine gender-specific words. The second model generates gender-specific words based on the appearance of persons in the scene. This model incorporates an equalizer, which ensures equal gender probability when gender evidence is occluded and confident predictions when gender evidence is present. Further, it limits gender evidence to the visual aspects of persons.

After considering the two models, students are asked if either or both might perpetuate or amplify gender biases and, if the answer is positive, whether these models may solidify harmful stereotypes. They are then asked to consider which demographic groups might be rendered vulnerable to harmful stereotypes as a result of using the software and how such vulnerability could be prevented.

Lessons Learned:

Student response to the module was positive when it was taught in the spring of 2019. In follow-up surveys, 85.1% of students reported that they found the module interesting. *77.7%* said that participating in the module helped them think more clearly about the ethical issues discussed. 85.1% said that the module increased their interest in learning about the ethical issues discussed.

A few things we learned from the experience:

The philosophical content and questions could be more strongly motivated. The module could begin with a more engaging activity so as to prevent passivity and make the ethical problem appear more urgent and engaging. Likewise, having more specific ethical questions on the table from early on could help frame and orient the exercise and encourage more in-depth philosophical discussion.

It would be ideal if a computer scientist could present the technical material, which is necessary for the module but requires some technical fluency to be made more interesting.

Technical terms and key philosophical questions should be explained at depth, and examples of abstract ideas should be given so as to maximize clarity and improve the quality of philosophical discussion.