Repository Entry
Embedded EthiCS @ Harvard Teaching Lab

Overview

Course:	CS 187: Introduction to Computational Linguistics and Natural-language Processing
Course Level:	Upper-level undergraduate
Course Description:	"Natural-language-processing applications are ubiquitous: Alexa can set a reminder if you ask; Google Translate can make emails readable across languages; Watson outplays world Jeopardy champions; Grover can generate fake news, and recognize it as well. How do such systems work? This course provides an introduction to the field of computational linguistics, the study of human language using the tools and techniques of computer science, with applications to a variety of natural-language-processing problems such as these. You will work with ideas from linguistics, statistical modeling, and machine learning, with emphasis on their application, limitations, and implications. The course is lab- and project-based, primarily in small teams, and culminates in the building and testing of a question-answering system."
Module Topic:	Free Speech and Content Moderation Online
Module Author:	Ellie Lasater-Guttmann
Semesters Taught:	Fall 2021
Tags:	GPT-3 [CS], Grover [CS], NLP [CS], free speech [phil], censorship [phil], misinformation [phil], harm [phil], J.S. Mill [phil]
Module Overview:	This lab-format module uses John Stuart Mill's arguments about free speech and censorship to evaluate the ethics of three types of cases: (1) posting content online, (2) human moderators censoring content online, and (3) using Grover (an automated content moderation tool) to remove posts labeled as AI-generated. Students label cases of speech or censorship as being ethically permissible, impermissible, or obligatory, with most cases concerning the spread of falsehoods or AI-generated speech. Examples were evaluated on two primary dimensions: (A) whether AI-generated content should be censored like human speech and (B) under what circumstances false information should be limited online.

Connection to Course Material:	Students have learned how to generate language using GPT-3 software and have grown familiar identifying instances of NLP-generated language.	Grover acts as an elegant entry into issues of free speech for the digital age, given that it can so easily be used for censorship. Grover is a compelling software tool that builds naturally upon content already discussed in the course. Free speech/censorship helpful tools for evaluating Grover's usefulness.

	Goals	
Module Goals:	- Discuss limits of free speech - Identify possible harms with spreading falsehoods - Apply ethical reasoning about speech and censorship from outside the tech domain to artificially generated speech - Identify cases where the principles governing free speech and censorship can come apart	
Key Philosophical Questions:	1. What type of content should be censored? 2. What are the unique ethical features of AI-generated speech? 3. When should falsehoods be protected speech? 4. How do we balance the harms of false-positives/false-negatives with the benefits of software like Grover?	See the Lessons Learned section on how these philosophical questions could have been sharpened.

	Materials	
Key Philosophical Concepts:	● Free speech ● Censorship ● The harm principle ● Misinformation ● Ethically impermissible, permissible, obligatory	To answer our first philosophical questions, students label different cases of censorship as being permissible, impermissible, or obligatory. Labeling individual cases allows them to brainstorm overarching principles that govern censorship.

Assigned Readings:	● J.S. Mill's "On Liberty" ch 2 (sections)	We considered Mill's arguments on (limitless) free speech and then (relatively limited) censorship and how they come apart. This reading should be obligatory before the module and provides a nice lens for students when contrasting their own views with Mill's. Mill's *On Liberty* paired well with the module material because it revealed that principles governing censorship and those governing free speech can come apart.

	Implementation	
Class Agenda:	1. Introduce concepts of ethically impermissible, obligatory, permissible 2. Discuss Mill's arguments about free speech 3. Activity Part 1: Posting on Social Media 4. Activity Part 2: Content Moderation 5. Grover introduction 6. Activity Part 3: Automating content moderation with Grover	
Sample Class Activity:	I developed a lab packet that was broken into three sections: posting on social media (speech), content moderation (censorship done on individual bases), and Grover (censorship done on an automated basis). Students labeled cases in each section as being ethically permissible, impermissible, or obligatory. What follows is the final section of the packet: **Part 3: Automation (~15 minutes)** Prompt: Answer the following questions in order. (1) Would it be ethically impermissible to implement Grover to identify posts	Because the class was conducted in a lab format, this exercise was certainly valuable for the class.

containing AI-generated speech and automatically remove them? Remember:

(a) This would remove weather reports like Post 2.
(b) Grover has a 2% false negative rate (AI-posts that will slip through the cracks).
(c) Also, Grover has a 2% false positive rate. This means that it would remove about 6.2M posts / day that were generated by humans in the USA.

Yes or No? What features of the case seem ethically relevant to your answer?

(2) Does your opinion about automatic-removal of posts depend on the content of those posts? **What types of posts can be automatically removed, ethically speaking? What types can't?**

(3) Would your opinion change if AI posts were posted 10,000 times / minute? Would you have the same opinion if a human were doing it (but at a rate of 1 posts / minute)?

(4) Assuming we could limit Grover according to anything you listed in question 2 of this section, would you still be comfortable using Grover if its false positive cases were all from a minority group (e.g. Grover removed posts from humans who posted in African-American Vernacular English - AAVE)? Would you want to make adjustments?

Module Assignment:	Students responded to the following essay prompt: Should AI-generated posts to social media services be moderated following the same guidelines as human-generated posts?	The assignment is intended to prompt students to reflect on whether AI-generated speech should be limited along the same dimensions as human-generated speech. With this more focused assignment, students can take a step back and review what they discussed in the module. The goal is to have students remind themselves of module content, rather than generate new material entirely.
Lessons Learned:	1. Putting the class activities into a lab packet was very successful. For a course that typically uses labs rather than lectures, I would strongly recommend a lab-based module where the learning is accomplished in small groups. 2. As written, students came to a variety of (compelling) conclusions about free speech, censorship, and automatic content moderation, but those conclusions were sufficiently varied that the points of reconvening the whole class were difficult to manage. Instead, I would change the packet slightly to focus on a single concern - e.g., misinformation. Have all of the examples be instances of misinformation spreading (particularly those using AI).	