CSCI 699: Ethics in Natural Language Processing

Units: 4

Fall 2023 TuTh 4-5:50pm

Location: VHE 206

Instructor: Jieyu Zhao

Office: TBD

Office Hours: TBD

Contact Info: jieyuz@usc.edu. I will reply to the email in 48 hours. Please make sure to include “699” in the email subject.

Course Description

Although there have been impressive advancements in natural language processing (NLP), several studies reported that NLP models contain social biases. Even worse, the models run the risk of further amplifying the stereotypes and causing harms to people. As NLP technology continues to advance and be integrated into various domains such as healthcare, finance, marketing, and social media, it raises important ethical concerns that need to be addressed. In this course, students will critically examine the ethical implications of NLP, including issues related to bias, fairness, privacy, transparency, accountability, and social impact. Through discussions, case studies, and guest lectures, students will explore the ethical challenges associated with NLP and develop a deep understanding of the ethical considerations that arise when designing, implementing, and deploying NLP applications.

Learning Objectives and Outcomes

Students will get a broad understanding about possible issues in current NLP models and how current research has tried to alleviate those issues. This class will equip students with the ability to read and write critical reviews about research papers. At the same time, they will learn how to conduct research related to NLP fairness, interpretability and robustness.

Recommended Preparation:

- Familiarity with natural language processing (at the level of CSCI 544) and machine learning (at the level of CSCI 567).
- Programming skills. We will mainly use python with PyTorch, but you can use any other libraries for your final project.

Course Notes

Grading type: Letter of Credit/No-Credit. Lecture slides will be posted online after the class.

Technological Proficiency and Hardware/Software Required

For the course project, access to computational resources (e.g., GPU) is highly recommended.

Required Readings and Supplementary Materials

All reading materials will be posted on the course website at the beginning of the course.

Supplementary materials

The following courses are relevant:

- UW: Linguistics 575: Ethics in NLP: http://faculty.washington.edu/ebender/2017_575/
- Berkeley: CS 294: Fairness in Machine Learning: https://fairmlclass.github.io/
- CMU: Computational Ethics for NLP: http://demo.clab.cs.cmu.edu/ethical_nlp2019/

Description and Assessment of Assignments

- Grading policy:
    - 60% Course Project
    - 30% Paper Presentation
    - 10% Attendance and discussion participation

Course Project (60%)

Each student needs to individually finish one research project related to the class topics. There should be a "deliverable" result out of the project, meaning that your project should be self-complete and reproducible (scientifically correct. A typical successful project could be: 1) a novel and sound solution to an interesting research problem, 2) correct and meaningful comparisons among baselines and existing approaches, 3) applying existing techniques to a new application. We will not penalize negative results, as long as your proposed approach is thoroughly explored and justified. Overall, the project should showcase the student's

ability to think critically, conduct rigorous research, and apply the concepts learned in the course to address a relevant problem in the field of NLP ethics.
Students should use the standard *ACL paper submission template to finish their writing report regarding the course project.

- **Project proposal (10%)**
  Students are expected to finish a 2-page long project proposal by Week 5. The proposal should articulate the research question, justify the significance of the research, and provide evidence of the student's knowledge and understanding of the research literature. A timeline for the project will be highly recommended to be included in the proposal.
- **Midterm progress report (10%)**
  By Week 10, the students should finish a ~3-page progress report. The report should provide a clear statement about the research goal (could be different from the original one), a concise overview of the work completed so far, including any challenges encountered and solutions implemented and a report of some initial results.
- **Final presentation (20%)**
  During the last two week, the students will make a 30-minute presentation about their project. It should include the research goal, the motivation, related work, their methodology and results. There will be a 5-minute QA session for other students to ask questions.
- **Final project report (20%)**
  Students will write a final project report to describe the details about their research. The report should follow the NLP conference paper format, including the abstract, introduction, related work, result demonstration and discussion section. If the result is negative, it won't be penalized but the students should highlight their analysis about what could be the possible reasons. The report should be in total 8 pages (excluding the reference).

Paper Presentation (30%)

- Paper presentation will help students to develop the skills to give research talk to others.
- Each student will present 2 papers to the class. The student will prepare the slides for the paper and lead the discussion.
- Each week, there will be another student signed up as the feedback provider (reviewer). The presenter should finish the rehearsal of their talk with the reviewer. The reviewer will finish a feedback form at least 2 days before the class.
- Reviewer will provide the feedback to the presenter. The instructor will grade the presentation. Grading rubrics: correctness of the content (40%), clarity (20%), discussion (20%), slides & presentation skills (20%).

Participation (10%)

Students are expected to attend the class and get involved in paper discussion. This includes asking questions about the presentations or express their opinions on the topics.

Grading Breakdown

Assignment	Points	% of Grade
Participation	10	10
Paper presentation	30	30
Project proposal	10	10
Project midterm progress report	10	10
Project final presentation	20	20
Project final report	20	20
TOTAL	100	100

Assignment Submission Policy

Assignment will be submitted to google drive by 11:59 pm on the due date.

Grading Timeline

Grading will happen within one week of submission.

Additional Policies

Students will have in total **4 late days** to use for the project proposal and progress report (no late days for the final report). The grace period will be used in integer amounts. Additional late days will result in a deduction of 10% of the grade on the corresponding assignment per day.

Course Schedule: A Weekly Breakdown

	Topics/Daily Activities	**Readings and Homework**	**Deliverable/ Due Dates**
Week 1	Course introduction; paper candidate list discussion; review about how to do research presentation	Presentation signup	By W2
Week 2	Philosophical Foundations		
Week 3	Social biases in NLU		
Week 4	Social biases in NLG		
Week 5	Causal View of Fairness		Project proposal due by 09/24/2023 11:59 pm
Week 6	Privacy and Security in NLP		
Week 7	Civility & Toxicity		
Week 8	Misinformation & Manipulation		
Week 9	Distributional Robustness in NLP		
Week 10	Language and Society		Project midterm progress report due by 11/29/2023 11:59pm
Week 11	Model Explanation in NLP		
Week 12	NLP for social good		
Week 13	Human-Centered AI		
Week 14	Final project presentation		
Week 15	Final project presentation		
FINAL	Final report		*Due on the university-scheduled date of the final exam.*

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, comprises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

Other violations of academic integrity include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), collusion, knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Students and Disability Accommodations:
USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Support Systems:
Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call

The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline is comprised of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.