CSI 269: Special Topic in AI: Fairness, Accountability, and Transparency in Natural Language Processing
Kai-Wei Chang

Alert: this seminar course is project-heavy and required to read many papers!

Course information:

- **Lectures**
  - Time: Tu/Th 4pm-5:50pm
  - Location: HAINES A18
- **Instructor**
  - Prof. Kai-Wei Chang
    - Office hour: Tu 3-4pm (tentatively)
    - Office location: Eng VI 374
- **Course Description:** Recent advances in data-driven machine learning techniques have revolutionized natural language processing (NLP) applications. Despite the remarkable accuracy of these approaches, learning algorithms run the risk of relying on societal biases encoded in the training data to make predictions. Therefore, these algorithms risk potentially encouraging unfair and discriminatory decision making and raise serious privacy concerns. Without properly quantifying and reducing the reliance on such correlations, the broad adoption of these models might have the undesirable effect of magnifying harmful stereotypes or implicit biases that rely on sensitive demographic attributes. In this course, we will examine real-world NLP applications and the underlying ethical issue. The course consists of paper reading, paper presentations, and a final project.
- **Course Activities:** The instructor will give lectures in the first two weeks. Then, the rest of the quarter will mostly consist of paper presentations and discussion. Students are expected to read the required reading before classes.
- **What you will learn:** You will learn the recent trends in building fair, accountable, transparent, and ethical NLP models as well as learn the skills for conducting NLP research by conducting a literature survey, giving a paper presentation, working on a final project, demonstrating your research work, and write critical reviews.
- **Tentative Topics:**
  - **Fairness in NLP.** Misrepresentation and Bias. Bias in NLP applications (coreference resolution, machine translation, recommendation system, etc).
- **Algorithmic fairness.** Definition of fairness. Calibration methods. Techniques for removing Bias.
- **Inclusive NLP.** NLP for low-resource language and social good applications.
- **Robustness in NLP.** Adversarial attack and data bias in NLP.
- **Explainable NLP.** Transparency and visualization of NLP Models.

**Prerequisites:**
- Machine learning (CSM146/CS260) or equivalent is required before taking this seminar course. We will not cover ML background in this course.
- Basic math skill
- Programming skill. We will mainly use python with PyTorch, but you can use any other libraries for your final project.

**Related Courses**
- List of related courses: [https://aclweb.org/aclwiki/Ethics_in_NLP](https://aclweb.org/aclwiki/Ethics_in_NLP)

  The following courses are especially relevant:
  - Berkeley: CS 294: Fairness in Machine Learning: [https://fairmlclass.github.io/](https://fairmlclass.github.io/)

**Course policy**
- **Grading policy:**
  - 40% Final Project
  - 20% Paper Presentation
  - 20% Literature Review
  - 10% feedback (5% presentation rehearsal feedback; 5% final project review)
  - 10% attendance/participation
- **Accommodations for students with disabilities:** To insure that disability-related concerns are properly addressed from the beginning, students with disabilities who require assistance to participate in this class are asked to see me as soon as possible. We aim to be as accommodating and fair as possible.
- **Use of outside resources:** When using outside resources, proper citation is necessary. This includes papers, text books, software libraries, websites, and helps from others. For the details of Honor Code, please refer to [UCLA Student Conduct](https://www.studentconduct.ucla.edu/). If you have any doubt, please check with me in advance. You may get F in the final letter grade if we detect any cheating.
- **Late policy:** The submission site will be closed 1hr after the deadline. No late submission will be accepted (unless emergency).
- **Collaboration:** By default, students in the same team will get the same score unless special circumstances. We encourage students to use a version control system (e.g.,
It is important to keep your hard work in a safe place and log the contributions of individuals. If your team members complain about you and you cannot provide evidence of your contribution, we may lower your score.

### Paper presentation

- The purpose of this project presentation is to allow students to practice giving research talks in front of others. Imagine that you’re presenting your work at a conference in front of people having a basic understanding of NLP.
- Each presentation is 30~50 min (TBD) and each group has at most 2 students.
- The instructors will provide a list of papers for presentation. Try to make your presentation simple and focus on the big ideas, but also provide enough details for understanding the core technique.
- **One week** before the presentation, the presenting group should upload their slides and schedule a rehearsal with another group (feedback group). The feedback group has to complete a feedback form **no later than 3 days** before the presentation. The presentation team will be graded based on their performance of presentation in class and the feedback team will be graded by the quality of their feedback.

### Literature Review

- The goal of the literature review is to learn how to conduct a literature survey.
- The review contains two parts
  - Identify a list of papers related to the survey topic.
  - Write a "blog post" to explain a paper in layperson language.

### Final project

- The goal of the course project is to provide students with an opportunity to explore research directions on a related topic. Therefore, the project aims at producing a “deliverable” result, meaning that your project should be self-complete, reproducible (scientifically correct) and related to the course content. A typical (successful) project consists of 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 the proposed approach is well explored. It is recommended to form a group with a diversity of background, but not required.
• The instructor will provide a list of potential final project topics. Students are welcome to propose other project ideas.

• Final project schedule:
  ○ Project proposal: each group should provide a 1-page project proposal at the end of week 2. The template will be provided.
  ○ Project midterm report: by week 8, each group should provide a 2-page midterm report (introduction/related work/methodology/proposed experiments)
  ○ The project presentation is scheduled in week 10. Each group is expected to prepare a 10~15 min presentation, and the slides are due 1 day before the presentation.
  ○ A 4-page final project report is due around the final week.
  ○ During the scheduled final exam time, we will have a demo session. Students are expected to present their work to each other and the public audience.
  ○ Each student will be assigned to review a project by others. The review is due before the final week. A review template will be provided.
  ○ The final project will be graded by the instructor, and detailed grading rubrics will be announced.

Tentative Schedule:

• Week 1: Introduction/overview (lecture) (presentation sign-up)
• Week 2: NLP and ML Background (lecture) (presentation sign-up)
• Week 3: Biases in Language Representation (student presentation starts)
• Week 4: Biases in NLP Models and How to Mitigate Them (literature survey due)
• Week 5: Algorithmic Fairness
• Week 6: Inclusive NLP
• Week 7: Robustness in NLP
• Week 8: Data Bias and Domain Adaptation (mid-term report due)
• Week 9: Explainable NLP
• Week 10: Final project presentation
• Final week: Demo session. The final report and review are due.