11-830 Computational Ethics for NLP

Lecture 2: Ethical Challenges in NLP

Using Human Subjects

Carnegie Mellon University
Language Technologies Institute

Human Subjects

- We are trying to model a human function
- Labels are certainly noisy
- How to use humans to find better labels/know if they are right
- Let’s put it on Amazon Turk and get the answer

History of using Human Subjects

- WWII Nazi and Japanese prisoners in concentration camps
  - Medical science did learn things
  - But even at the time this was not considered acceptable
- Tuskegee Syphilis Experiments
- Stanford Prison Experiment
- Milgram experiment
- National Research Act of 1974

Tuskegee Syphilis Experiment

- Understand how untreated syphilis develops
- US Public Health System 1932-1972
- Rural African-American sharecroppers, Macon Co, Alabama
  - 399 already had syphilis
  - 201 not infected
- Given free health care, meals and burial service
- Not provided with penicillin when it would have helped
  - (Though not known at the start of the experiment)
- Peter Buxton, whistleblower, 1972

Doctor taking blood from Tuskegee Subject
[National Archives via Wikipedia]

Stanford Prison Experiment

- Philip Zimbardo, Stanford University, August 1971
- Test how perceived power affects subjects
- Groups arbitrarily split in two
  - One group were defined “prisoners”
  - One group were defined “guards”
- “Guards” selected uniforms, and defined discipline

https://www.youtube.com/watch?v=oAX9b7agT9o

Blue vs Brown Eye “Racism”

- Kids separated by color of eyes
  - Blue eyes are better
  - Brown eyes are worse
- Quickly separate in clans
- Blue given advantages, Brown given disadvantages
- Kids quickly live our the divisions
- Is this experiment ethical?
- Do we learn something
- Do the participants learn something?

_x0001_ https://www.youtube.com/watch?v=KHxFuO2Nk-0

Milgram Obedience Experiment

- Stanley Milgram, Yale, 1962
- Three roles in each experiment
  - Experimenter
  - Teacher (actual subject)
  - Learner
- Learner and Experimenter were in on the experiment
  - Teacher asked to give mild electric shocks to the Learner
  - Learner had to answer questions and got things wrong
  - Experimenter, matter of factly, asked Teacher to torture Learner
- Most Teachers obeyed the Experimenter

Ethics in Human Subject Use

- These experiments (especially the Tuskegee Experiment)
- Led to the National Research Act 1974
  - Requiring “Informed Consent” from participants
  - Requiring external review of experiments
  - For all federal funded experiments

IRB (Ethical Review Board)

- Institutional Review Board
  - Internal to institution
  - Independent of researcher
- Reviews all human experimentation
  - Assesses instructions
  - Compensation
  - Contribution of research
  - Value to the participant
  - Protection of privacy

IRB (Ethical Review Board)

- Different standards for different institutions
  - Medical School vs Engineering School
- Board consists of (primarily) non-expert peers
- At educational institutions also
  - Help education new researchers
  - Make suggestions to find solutions to ethics problems
    - How to get informed consent on an Android App
    - “click here to accept terms and conditions”

Ethical Questions

- Can you lie to a human subject?
- Can you harm a human subject?
- Can you mislead a human subject?

Ethical Questions

- Can you lie to a human subject?
- Can you harm a human subject?
- Can you mislead a human subject?

- What about Wizard of Oz experiments?
- What about gold standard data?

Using Human Subjects

- But its not all these extremes
- Your human subjects are biased
- Your selection of them is biased
- Your tests are biased too

Human Subject Selection Example

- For speech synthesis evaluation
  - Listen to these and say which you prefer
- Who do you get to listen
  - Experts are biased, non-experts are biased
- Hardware makes a difference
  - Expensive headphones give different result
- Experiment itself makes a difference
  - Listening in quiet office vs on the bus
- Hearing ability makes a difference
  - Young vs old

Human Subject Selection

- All subject pools will have bias
  - So identify the biases (as best you can)
  - Does the bias affect your result (maybe not)
- Can you recruit others to reduce bias
  - Can you do this post experiment
- Most Psych experiments use undergrads
  - Undergrads do experiments for course credit

Human Subject Selection

- Most IRB have special requirements for involving
  - Minors, pregnant women, disabled

Human Subject Selection

- Most IRB have special requirements for involving
  - Minors, pregnant women, disabled
- So most experiments exclude these
- Protected or hard to access groups are underrepresented

Human Subject Research

- US Government CITI Human Subject Research
  - Short course for certificate
- All Federal Funded Projects **require** HSR certification
  - You should do it NOW.
- Most IRB approval require CITI certification
  - You should do it NOW

We'll Use Amazon Mechanical Turk

- But what is the distribution of Turkers
  - Random people who get paid a little to do random tasks
- Its a large pool so biases cancel out
  - There are maybe 1000 regular highly rated workers
- Can you find out the distribution?
  - Maybe, but the replies might not be truthful
- Does it matter?
  - Depends, but you should admit it

Real vs Paid Participants

- Paying people to do use your system
  - Not the same as them actually using it.
- Spoken Dialog Systems (Ai et al. 2007)
  - Paid users have better completion rates
  - ASR word error rate different paid vs real (Black et al. 2011)
  - Paid, happy to go to wrong place (DARPA Communicator 2000)
    - User: “A flight to San Jose please”
    - System: “Okay, I have a flight to San Diego”
    - User: “Okay”
    - :-(

Human Subjects

- Unchecked human experimentation
- Led to IRB reviews of human experimentation
- All human experimentation includes bias
  - Admit it, and try to ameliorate it
  - Is your group the right group anyway
  - Experimentation vs Actual is different