Collecting Data from Human Subjects: Social Media & Crowdsourcing

Ethics in NLP: Including Society in Discourse and Design

February 14th, 2019

Ryan Georgi

Two Sides of Data Collection: Old Paradigms

- **Generating Data**
  - Annotators (employees)
  - Members of the public (human subjects)
- **Using Existing Data**
  - Previously generated data
  - Text/speech published in print or on tape

Two Sides of Data Collection: Shifting Paradigms

- **Generating Data**
  - Paid — Crowdsourced Workers (e.g. Amazon Mechanical Turk, Crowdflower)
  - Unpaid — Volunteer Editors (e.g. Wikipedia, Amazon Product Reviews)
- **Using Existing Data**
  - Social Media
  - Scraping Web Sites
    - Text, Speech, Video, Images…

Data Collection:
Generating Data

- If you want someone to help you by labelling language data, are they:
  - An employee?
    - If so, fine, but you must abide by labor laws.
  - A member of the public?
    - In that case, the DHHS definition of IRB-covered research includes surveys
      - Personal opinions are considered ***private information***.
      - Answers to questions about things, products, or policies are ***exempt***.
      - Annotation is ***typically*** exempt, but you should make sure to be cautious.
    - As Gilles et al (2014) point out, human subjects can be compensated
      - …but care should be taken that compensation isn't coercive

Crowdsourcing/Microworking:
Labor Considerations

- With AMT, workers are nominally independent contractors
  - The pay rate is usually far below US minimum wage (not all workers are US-based)
  - Contract is defined by Amazon, not negotiated with either workers or data consumers

Crowdsourcing/Microworking:
Data Considerations

- Low pay/reward
  - Can mean low quality work
- Data provenance
  - What are the demographics of the crowdsourced workers?
    - Important if you ask questions that may have sociological components
  - What are their language skills?
    - They may be a native English speaker, but not a native *American* English speaker

Data Collection:
Using Existing Data

- Disruptions to old paradigm, via Halford (2018)
  - Data have already been created
  - Data are not in our ownership
  - Data are not finite

Using Existing Data:
Data Have Already Been Created

- **Previously, NLP research was done on**:
  - Published data (Newswire, public-domain literature, etc)
  - Corpora produced under an IRB or employer/employee paradigm
- **Social media data, web-scraped data**:
  - No oversight from the researchers
  - No guidelines for "annotation"
  - Often no insight into provenance of data

Using Existing Data:
Data are not in our Ownership

- **Previously**:
  - Data collection in an IRB setting — obligation to anonymize, safeguard
- **Social media data, web-scraped data**:
  - Researchers don't have any control over the data storage.
  - All data is owned/controlled by the company whose platform it is on
  - …researchers can't make previously assumed guarantees

Using Existing Data:
Data are not Finite

- **Previously**:
  - Data collection was ***scoped***, performed, finalized, stored.
  - An important IRB exemption — data is ***not generalizable*** or ***not private***
  - Right to withdraw was easily done by opting out prior to finalization of project
- **Social media data, web-scraped data**:
  - Have no guarantees over what content is included
  - Right to withdraw theoretically exists (right to be forgotten, GDPR, etc)
    - …but is it followed in practice?

Reading Questions:

- What kind of "disruption" to the methods of data collection does the paper address? If one is not directly addressed, come up with an example related to the concerns of the paper.
  - (e.g.: "publishing" your words on Twitter vs. in 1970; Informed consent vs. ToS)
- Are there any specific issues that you encountered in the paper that are relevant to the principles of the Belmont Report (Links to an external site.)Links to an external site.? (Respect for Persons, Beneficience, Justice) Which ones?
  - (e.g. Does this data collection potentially focus on a protected group? A coerced one?)
- Give one suggestion of a "best practice" or rule of thumb that might be used to improve the data collection in the domain discussed in the paper.