The Social Impact of Natural Language Processing

- Research/development of NLP tools has traditionally not required the same rigour of ethical approval as similar scientific research (e.g. medical sciences/clinical work)
- This is due to the (misconception) that NLP and language data does not typically directly involve human subjects
- However, the growth of NLP makes clear that an ethical framework must exist to research/build these tools as NLP can increasingly have direct effects on people
- Most initial work focused on the data privacy component of NLP but this paper focuses on the effects of NLP technology on social justice
- NLP technology directly affects individuals
  - increased use of social media data can lead to indentifiability issues of users
  - language can be an indirect proxy for social identity/protected classes
  - detection of these characteristics through language and building models that use these characteristics will lead to biased/discriminatory models
- Some specific categories of social impact are discussed:
  - **Exclusion**
    - datasets typically carry demographic bias
    - models can overfit to demographic bias and not generalize to/misrepresent underrepresented groups in the data
  - **Overgeneralization**
    - the risk of incorrect models is compounded when the models deal with particularly sensitive topics where the impact of misclassification/misprediction are high (e.g. predicting race/gender)
    - we should not generalize/accept results from a model known to make costly mistakes when we may prefer to have no answer at all
  - **Topic Overexposure**
    - the availability heuristic contributes to confirmation bias where information that is readily available are taken to be more likely to be true or important
    - continued research in a topic can reinforce discrimination/bias towards the subjects of research
    - overexposure also leads to underexposure of other equally valid paths of study
  - **Dual Use**
    - NLP tools that are ostensibly used for good, can often be equally applied in unethical use-cases
- Overall, we need to build and apply our NLP tools with an understanding that we need to think about the ethical and social impacts of the technologies we create and the biases embedded within them

Principled Frameworks for Evaluating Ethics in NLP Systems

- Much work has been done in understanding the ethics of NLP systems and mitigating the negative effects of these tools/research
- However, there needs to be an understanding of what basis we should even decide whether or not things are ethical (borrowed from philosophy)
- Two frameworks are proposed:
  - **Generalization principle**
    - An ethical decision-maker must be rational in believing that the reasons for action are consistent with the assumption that everyone with the same reasons will take the same action
    - the evidence given to a decision-maker must not use more information than necessary and should only include task-relevant data
    - in practice, NLP must be interpretable and generalizable, and have well-defined behaviour, rely strictly on task-relevant information

  - **Utilitarian principle**
    * An action is ethical only if it is not irrational for the agent to believe that no other action results in greater expected utility
    * in practice, this shifts the focus on testing/evaluation NLP to mirror reality so we can analyze our models on how they will benefit/harm communities in the real world
  - these two frameworks can conflict for example, with the addition of demographic data for classification
    * generalization principle rejects this because we are adding demographic data that is not strictly relevant to the classification task
    * utilitarian principle accepts this because adding demographic data leads to a strictly more accurate system that benefits more people

Social Biases in NLP Models as Barriers for Persons with Disabilities

- NLP tools have inherited many of the biases against disabled people that exist in the real world
- text classification models are analyzed by perturbing text (replacing pronouns in sentences with references to persons with disabilities)
- the difference in output between the baseline and perturbed text gives a measure of how much reference to disability affects a model's predictions
- text containing references to disability (regardless of whether the term is recommended/non-recommended by disability rights organizations) often scores higher on toxicity and lower on sentiment than baseline text
- these biases can be significantly attributed to bias in the datasets that language models are trained on, as online discussions of disability is overwhelmingly in a negative context
- this has the effect of potentially suppressing neutral/positive discussions about disability, leading to less representation of views from those with disabilities (compounding the biased data problem!)