CS 1671 / CS 2071
Human Language Technologies

Session 26: Social factors, bias and ethics in NLP

Michael Miller Yoder

November 19, 2026



School of Computing and Information

Planning ahead after Thanksgiving break

- Guest lecture for next class, Tue Dec 1:
  - Luke Breitfeller, NLP in biomedical applications
- Class on Thu Dec 3 will be in-class project work time
  - Michael will be available to answer questions

Assessments: exam

- Exam grades have been released
- Email Yuelyu and Michael or use the Gradescope “regrade request” option if you have any issues
- Take a look at the class grading scale (≥92.5% is an A in the course)

Assessments: project

- Project final report is **due Tue Dec 8**
  - Implement your proposed LLM system, compare with your baseline system
  - See project website for instructions of what to include
  - Maximum 8-page report in ACL format (Word and LaTeX templates here)

Assessments: project

- Project presentation is **in final period**

  Thu Dec 10, 2-3:50pm

  - 5 points for participation
  - Michael will post a shared PowerPoint

Overview

- Language in social context
- NLP for computational social science
- Project example: classifier and dataset of white supremacist language
- Bias and ethics in NLP
  - *Warning: slides contain offensive stereotypes*

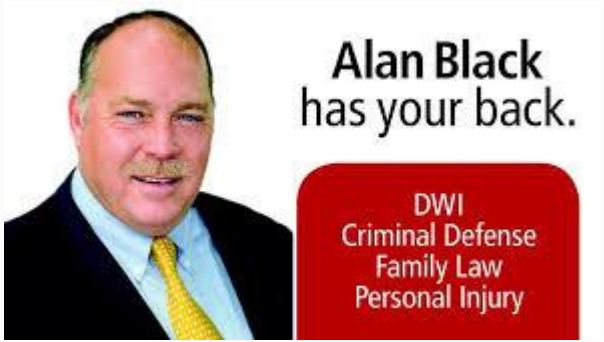
Language is embedded in social context

What different types of social situations do you encounter language in throughout your day?

What types of social contexts?

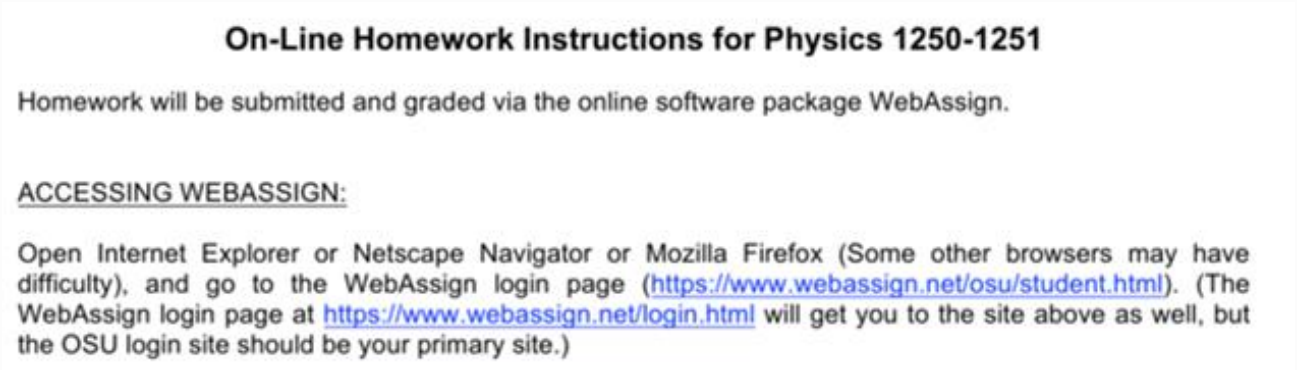






What types of social contexts?

Euro quals	2:45 PM ET ESPN3 BEL CYP	2:45 PM ET ESPN3 NED EST	2:45 PM ET ESPN2/ESPN3 WAL HUN	2:45 PM ET ESPN3 GER NIR	2:45 PM ET ESPN3 POL SVN



On-Line Homework Instructions for Physics 1250-1251

Homework will be submitted and graded via the online software package WebAssign.

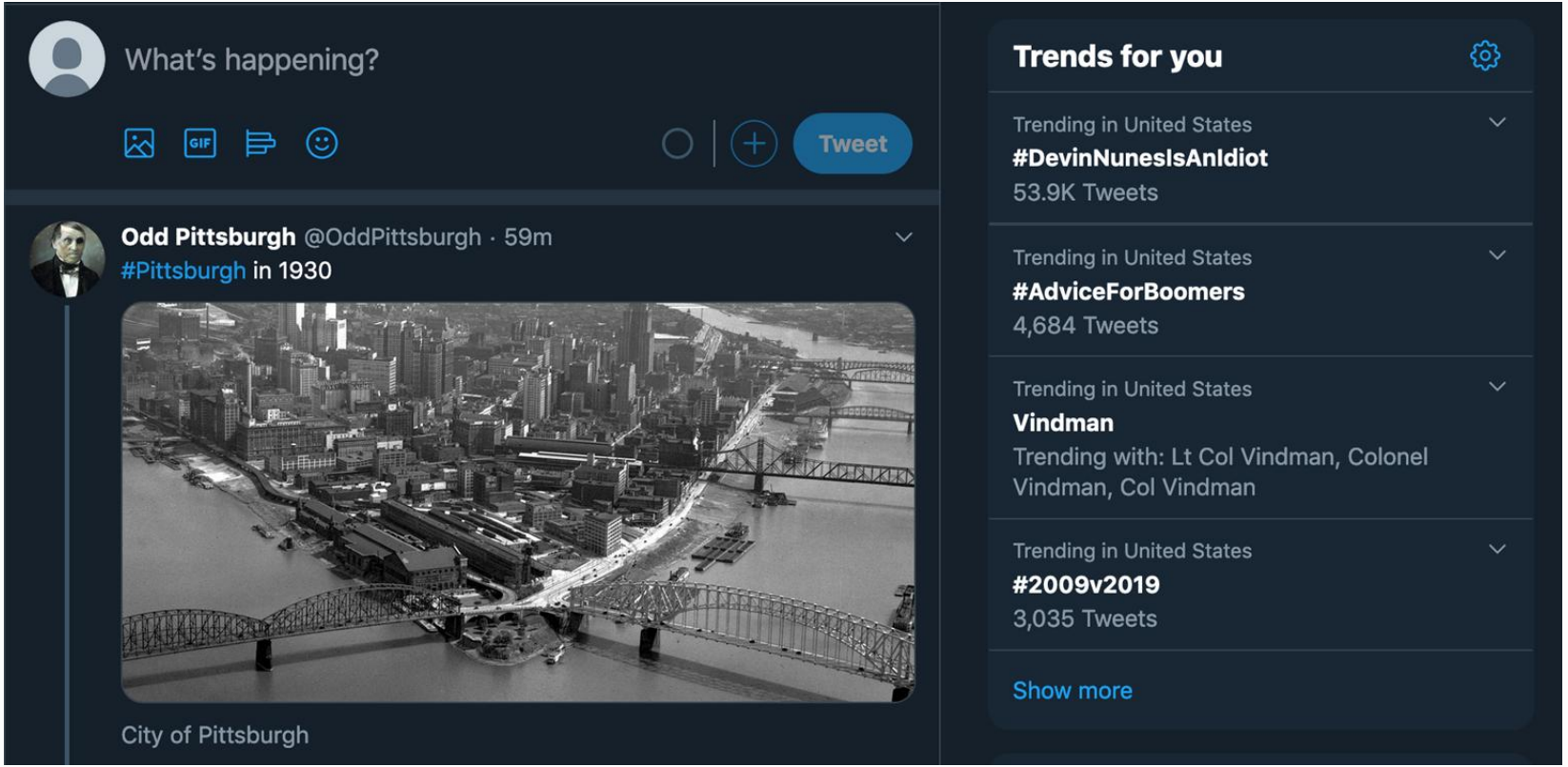
ACCESSING WEBASSIGN:

Open Internet Explorer or Netscape Navigator or Mozilla Firefox (Some other browsers may have difficulty), and go to the WebAssign login page (https://www.webassign.net/osu/student.html). (The WebAssign login page at https://www.webassign.net/login.html will get you to the site above as well, but the OSU login site should be your primary site.)

What types of social contexts?



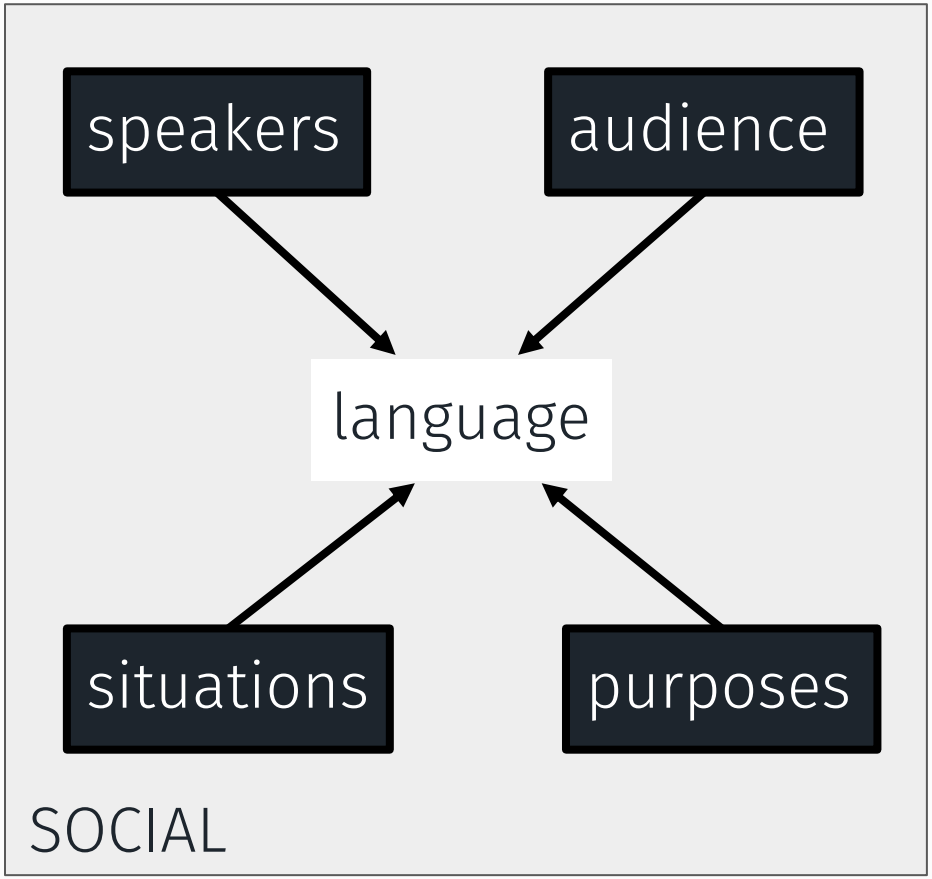
What types of social contexts?



What types of social contexts?






speakers
audience
language
situations
purposes
SOCIAL

NLP + social science: applications

hate speech detection

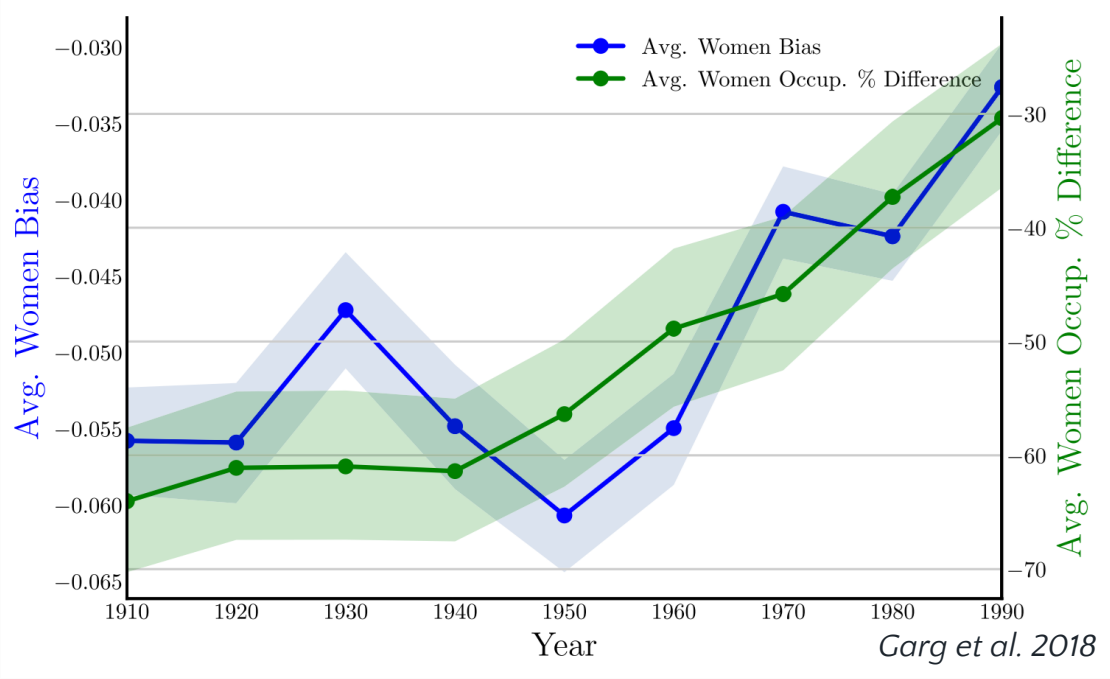


community norms



NLP + social science: applications

fairness and bias



Garg et al. 2018

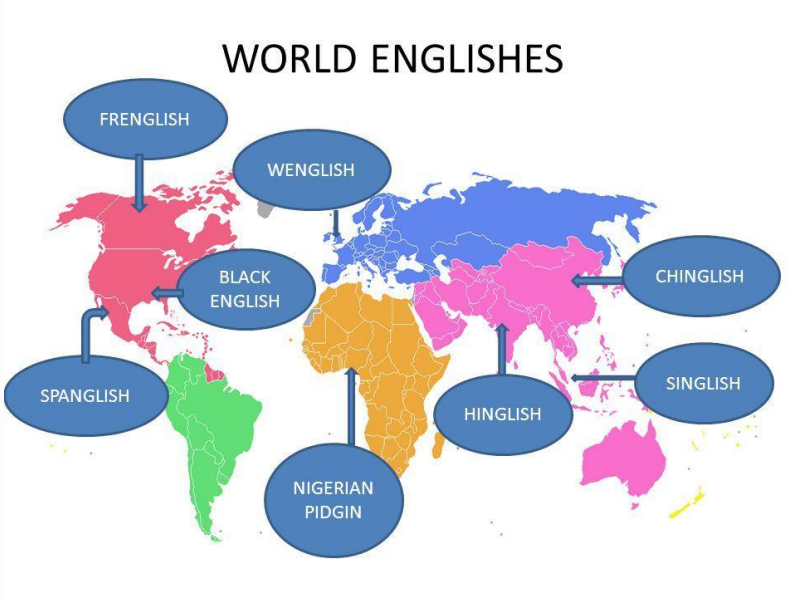
media framing



https://criticalmediareview.wordpress.com/2015/10/19/what-is-media-framing/

NLP + social science: applications

dialectal NLP tools



www.tes.com

Computational social science

Computational social science

- Investigating (modeling, analyzing) social phenomena with computational tools [Cioffi-Revilla 2017]
- CSS goal: find out something about **people** (social science)
- NLP goal: build computational tools that can process or produce language
- CSS+NLP: using NLP tools to measure or predict social information from language use

Computational social science: methods and data

- Observational studies, not lab or survey studies



large datasets of social interaction

Computational social science example

Example: How fast does fake news spread? [Vosoughi et al. 2018]

$$y = f(x)$$

spread through a network

true/fake news

network analysis

NLP/text mining

Computational social science example

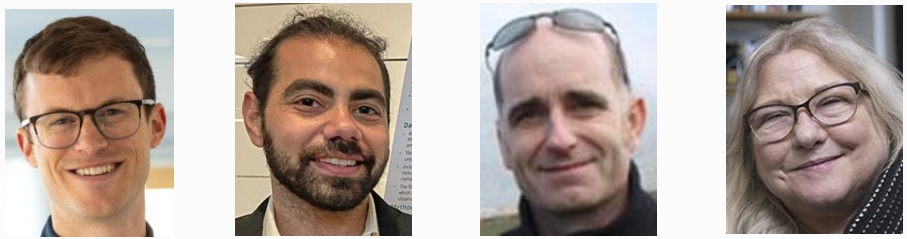
Example: Do police officers speak more respectfully to white drivers than Black drivers in traffic stops? [Voigt et al. 2017]

$$y = f(x)$$

social factors
(race of a driver)

language features
(politeness of officer talk)

Project example: A classifier and dataset of white supremacist language



Michael Miller Yoder, Ahmad Diab, David Brown, and Kathleen Carley. A Weakly Supervised Classifier and Dataset of White Supremacist Language. In *ACL (Volume 2: Short Papers)*. 2023

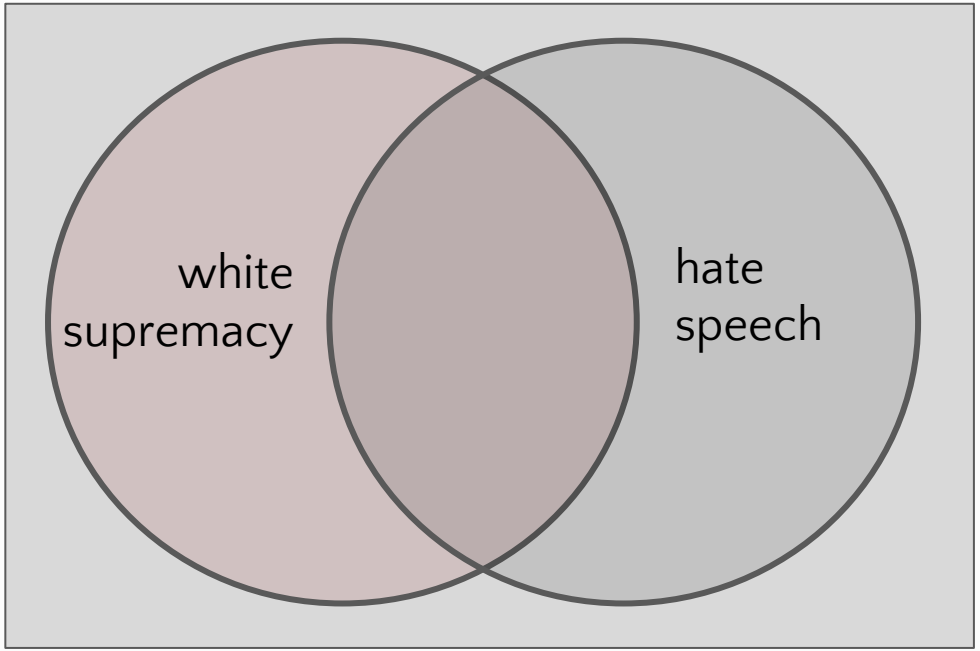
White supremacist extremism



Credit: Rolling Stone

Social movements advocating white male superiority and action against non-white peoples [Daniels 2009]

White supremacist language



"You should be fine posting about preserving the white race as long as you don't degrade other races."

Stormfront user

White supremacist dataset

- Text from discourse spaces or producers known for white supremacy
- 4.3M posts, 230M words
- 2001-2019

White supremacist

Forum posts: *Stormfront , 4chan [Papasavva et al. 2020, Jokubauskaitė & Peeters 2020], etc*

Tweets: *From hate groups [Qian et al. 2018], etc*

Articles: *Daily Stormer [Calderón et al 2021], manifestos [Pruden et al. 2022], etc*

Chat posts: *Patriot Front Discord*

White supremacist language classifier

- Paired white supremacist data with "neutral" and anti-racist data
- Anti-racist texts: a source of positive mentions of marginalized identities
- Finetune DistilBERT text classification models [Sanh et al. 2019]

Neutral

Forum posts: *r/politics, r/Europe, etc*

Tweets: *Matching year, shared vocab*

Articles: *Random US news articles*

Chat posts: *Random Discord dataset*

Anti-racist

Forum posts: *anti-racist subreddits (r/racism, r/BlackLivesMatter, r/StopAntiAsianRacism, etc)*

Tweets: *From anti-racist activists and groups (UNC Diversity & Inclusion list)*

Articles: *Medium posts tagged "anti-racism", "white supremacy", "racism", "BlackLivesMatter"*

Chat posts: *No source*

Model	Eval dataset			
	Alatawi et al. 2021	Rieger et al. 2021	Siegel et al. 2021	Mean
Siegel et al. 2021	60.3	61.8	61.3	61.2
Alatawi et al. 2021	**74.0**	81.2	89.7	81.6
DistilBERT annotated	65.3	86.1	92.9	81.4
DistilBERT weakly supervised	71.6	87.8	90.3	83.2
DistilBERT weakly supervised + annotated	70.9	**90.3**	**96.8**	**86.0**

ROC AUC scores of models (rows) on 30% test splits of evaluation datasets (columns)

Bias and ethics in NLP

Bias and ethics in NLP [Hovy and Spruit 2016]

- Language, society and individual are interrelated
  - We must think about ethics when dealing with people
- Demographic bias: language contains latent information about the people who produced it
  - Exclusion of people whose languages aren't represented in training data
  - Bias toward Indo-European languages and a few "high-resource" languages in NLP
- Dual use: reinforce prescriptive linguistic norms and degrade non-standard language use with educational language technologies "correcting" language
  - NLP can detect fake news, but also generate it
  - Authorship attribution could identify political dissenters
- Funding sources for research include militaries and governments. Whose interests are embedded in systems?

Allocational harms [Crawford 2017]

System unjustly allocates or withholds opportunities or resources to groups



Source: Gizmodo

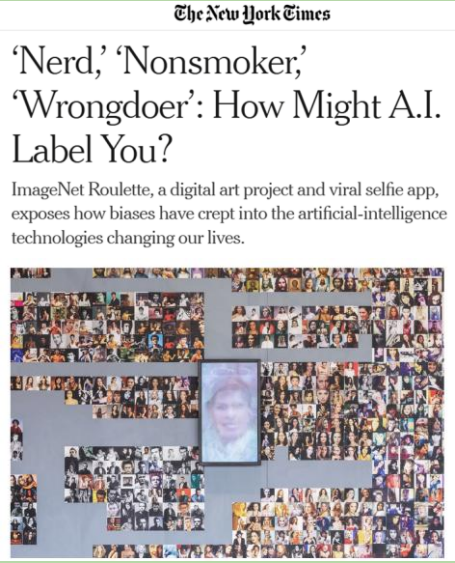


Representational (associative) harms [Blodgett et al. 2020]

System reinforces subordination or stereotypes about groups



Gender bias in word embeddings [Bolukbasi et al. 2016]


The New York Times

‘Nerd,’ ‘Nonsmoker,’ ‘Wrongdoer’: How Might A.I. Label You?

ImageNet Roulette, a digital art project and viral selfie app, exposes how biases have crept into the artificial-intelligence technologies changing our lives.

Crawford 2021

Bias & discrimination (representational harms)

Neural models encode social biases against marginalized identities

- Case study: language model generation.
- GPT-2 generates text with more **negative associations of black, woman, and gay demographics** on 'occupation' related topics.

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Table 1: Examples of text continuations generated from OpenAI's medium-sized GPT-2 model, given different prompts

Slide adapted from WebConf 2022 Tutorial on Societal Harms of LLMs.

The Woman Worked as a Babysitter: On Biases in Language Generation (Sheng et.al, 2019)

Causes of social harms from LLMs



> Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.
> – Ruha Benjamin, Race after Technology

- Language models were designed to model probability distributions of text. They
  - Do not understand social norms and morals
  - Reinforce and amplify biases
- Uncurated sources of training data
  - Reddit: 67% of Reddit users in the United States are men, and 64% between ages 18 and 29
  - Wikipedia: only 8.8–15% are from women editors
  - Web data contains conspiracy theories, misinformation, aggressive text

Slide adapted from WebConf 2022 Tutorial on Societal Harms of LLMs.

Discussion

- Can you think of other example of **allocational harms** unjustly distributing opportunities and resources, from NLP or other ML systems?
- What about **representational harms** that reinforce stereotypes about minoritized groups?

Power and structural issues in NLP ethics

It’s about power: ethical concerns of software engineers [Widder et al. 2023]

- Surveyed 115 software engineers and interviewed 21 software engineers about their ethical concerns, what happens when they develop ethical concerns and **what affects their power to resolve their concerns**
- Military, privacy, advertising, surveillance were top ethical concerns
- Scope of concern: from bugs to questioning entire purpose of an industry
- Refusal, ‘quiet quitting’ of productivity on a project
- Not continuing with interviews
- Seeking reassignment to other projects or trying to change the project
  - Pivoting from facial identification to facial verification (i.e. are these the same people?)
- Financial and immigration precarity can make doing something about ethical concerns difficult
- Organizational incentives (making $$) that might lead to ethical tradeoff, like selling user data to advertisers when scrambling to find a new revenue stream

Widder et al. 2023 implications

- Not so much about identifying issues, but giving programmers power to address them
  - Ethics checklists and codes can help empower individuals to do this
- Workers need ‘guidance on how to navigate organizational power dynamics’
- From a focus on good design -> critique of whose goals are being achieved
- Collective action often needed, but also the role of individuals refusing to work on projects
  - Are they replaceable, as Palantir treated them when employees left over ethical concerns over selling tech to US border enforcement (ICE)?

Language (technology) is power [Blodgett et al. 2020]

- Recommendations for better work on bias in NLP
- Look at fields outside CS for guidance
- Treat representational harms as harmful in their own right
- Explicitly state why “bias” in systems is harmful, in what ways, and to whom. Be explicit about normative reasoning behind these judgements
- Engage with the lived experiences of members of communities affected by NLP systems. Reimagine power relations between technologists and such communities.

Discussion

- Could you relate to any of the ethical concerns raised by software engineers?
- Have you ever had ethical concerns with any work you have done? If not, what types of work might you have ethical concerns about?

Conclusion

- Language is embedded in social context
- Computational social science studies people and societies with computational models of observational data
  - Often uses NLP for analyzing text
- Social biases can be encoded across the NLP system pipeline: data, modeling and use of systems
- Beyond design interventions, consider whose interests NLP systems are serving and push for greater accountability of such systems to all users