CS 384: Ethical and Social Issues in NLP

Dan Jurafsky
Stanford University
Spring 2020

Introduction and Course Overview

Thanks to Tsvetkov and Black course for ideas and slides!
How should we use NLP for good and not for bad?
The common misconception is that language has to do with *words* and what they mean. It doesn’t. It has to do with *people* and what *they* mean.

Herbert H. Clark & Michael F. Schober, 1992

Decisions we make about our data, methods, and tools are tied up with their impact on people and societies.
Hypothetical case

Should we use NLP to build IQ tests that determine student's IQ from the text they post on social media (or the text they write in school exams).

Intelligence Quotient: a number used to express the apparent relative intelligence of a person
IQ Classifier

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Our test results show 90% accuracy
  - White males have 95% accuracy
  - People with brown hair under age of 25 have only 60% accuracy
- Who is responsible?
  - Researcher? Reviewer? University? Society
IQ classifier

IQ tests are known to be racially and socio-economic status (SES)-biased. NLP systems likely to pick up on spurious correlations between intelligence metrics and linguistic features of racial or SES groups.
Hypothetical case

Should we use NLP to build a BERT-based neural detector for sexual orientation from social media text?
Sexual Orientation Classifier

Who can be harmed by such a classifier?

- In many countries being gay is prosecutable
- It might affect people’s employment; family relationships; health care opportunities;
- Personal attributes, e.g. gender, race, sexual orientation, religion are social constructs. They can change over time. They can be non-binary.
  They are private, intimate, often not visible publicly.
- these are properties for which people are often discriminated against.
Sexual Orientation Classifier

Where does the data come from?
Who gave consent?
Is the classifier interpretable?
These are easier cases

(Although they are both based on real research papers)

Most cases are more complex
Even earlier

Ethical questions have been part of NLP since the beginning
Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED

...

WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?
My father
YOUR FATHER
Ethical implications of ELIZA

People became deeply emotionally involved with the program

Weizenbaum's secretary asked him to leave the room when she talked with ELIZA

When he suggested that he might want to store all the ELIZA conversations for later analysis, people immediately pointed out the privacy implications

- Suggesting that they were having quite private conversations with ELIZA
“Hey, new question,” Barbie said. “Do you have any sisters?”

“Yeah,” Tiara said. “I only have one.”

“What’s something nice that your sister does for you?” Barbie asked.

“She does nothing nice to me,” Tiara said tensely.

Barbie forged ahead. “Well, what is the last nice thing your sister did?”

“She helped me with my project — and then she destroyed it.”

“Oh, yeah, tell me more!” Barbie said, oblivious to Tiara’s unhappiness.

“That’s it, Barbie,” Tiara said.

“Have you told your sister lately how cool she is?”

“No. She is not cool,” Tiara said, gritting her teeth.

“You never know, she might appreciate hearing it,” Barbie said.
What questions should we ask ourselves as we develop NLP technology?
One set of guiding principles: The Belmont Report

1. Respect for Persons
   ◦ Individuals as autonomous agents

2. Beneficence
   ◦ Do no harm

3. Justice
   ◦ Who should receive benefits of research and bear its burdens?
One set of guiding principles: The Belmont Report

Respect for Persons
  ◦ Are we respecting the autonomy of the humans in the research (authors, labelers, other participants)?

Beneficence: Do no Harm
  ◦ Who could be harmed? By data or by prediction errors?

Justice
  ◦ Is the training data representative?
  ◦ Does the system optimize for the “right” objective?
  ◦ What are confounding variables?
Who should decide?

- The researcher/developer?
- The user of the technology?
- Paper reviewers?
- The IRB? The University?
- Society as a whole?

We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences
Welcome to CS384!

Dan Jurafsky

Peter Henderson

Hang Jiang
Our goal

Survey NLP areas that deal with people
Where NLP has the potential to do harm or do good
And any of:
- Understand the ethical and social implications
- Build better systems
- Offer new ways of thinking
The duality of CS384

Do no harm

Do good
Cs384.Stanford.edu
Final Projects
Questions to Consider in Choosing a Topic

- **Structured:**
  - Task and data sets are well defined, can make rapid progress with existing NLP models
  - Work will likely not result in publication (maybe suitable for workshop venues) → though it depends how good your model is!

- **Semi-structured**

- **Unstructured**
Questions to Consider in Choosing a Topic

- **Structured**
- **Semi-structured:**
  - Some prior work on task. Data exists but may not be well-formatted or easy to approach
  - Research questions are clear but exact formulation of task is not
  - Project will require creativity in structuring tasks and may result in publishable work
- **Unstructured**
Questions to Consider in Choosing a Topic

- **Structured**
- **Semi-structured**
- **Unstructured**:  
  - Topic may be interesting, but research questions are unclear and hard to define  
  - Not clear what the correct data set is, may need to create one  
  - Could result in really great work, but will require substantial student effort (High risk high reward!)
Sample project intuitions in 3 areas

Drawn from Tsvetkov and Black course
Bias and Objectivity (Semi-structured)

- Lots of “challenge” data sets designed to identify social biases in models

- Choose a state-of-the-art NLP model:
  - Evaluate the model for bias on a challenge data set
  - Reduce the bias of the model: (data balancing, architecture changes, adversarial training objectives, etc.

- Possible data sets:
  - Coreference resolution: WinoBias, Winogender
  - Machine translation: WinoMT
  - Hate speech classification: [need to infer demographic labels]
Bias and Objectivity (Unstructured)

- Measure and/or mitigate bias in word representations
  - Word embeddings (Bolukbasi et al. 2016)
  - Contextualized word embeddings: ELMo, BERT (Kurita et al. 2019, May et al. 2019, Zhao et al. 2019)
  - Think about training data changes, architecture changes, adversarial training objectives, etc.

- Identify and quantify bias in domains and corpora
  - Computational social science: analyze text to measure bias in a community
  - Examples: online fiction writing, Wikipedia, economics job market forum
  - Linguistic cues of biased language, e.g. Wikipedia
  - Linguistic cues of bias across languages
Bias and Objectivity (Structured)

- Reimplement published methods and measure bias across several existing datasets
  - across languages
- Write a survey paper on Bias in NLP models and datasets
Civility in Communication (Structured)

- Develop a classifier to identify offensive/hate speech
- Lots of existing data sets: e.g. Davidson 2017, SemEval 2019 Task 5
- Any project on offensive language should address the risk of racial bias, but this does not necessarily need to be a focus of the task
Civility in Communication (Semi-structured)

- Offensive language in-context, e.g. forecasting derailment of conversation (e.g. Cornell toolkit for conversation analysis), or collecting new datasets of toxic/offensive/hate speech in context (e.g. from Twitter or Reddit)
- Identifying toxicity against Open Source developers (ping us for data)
Civility in Communication (Unstructured)

- Develop typologies of uncivil communication
  - Building on Breitfeller et al. 2019 or Wang&Potts 2019, collect more data or build classifiers to detect microaggressions

- Analyze impact of hate speech
  - Who does hate speech target? (Silva et al. 2016)
  - Audit shared workshop data (such as the SemEval 2019 task) to see who are most commonly the targets of hate speech in datasets (and who might be missing)
The Language of Manipulation (Structured)

- Develop a classifier to identify propaganda or fake news, using existing standard data sets:
  - Propaganda detection: [SemEval 2020](#) task on propaganda detection; also [NLP4IF 2019](#)
  - Fake news data set: [Perez-Rosas 2018](#)
  - Relation between headlines and main article text: [Fake news challenge](#)

- Develop a model to do fact-checking with existing standard data sets:
  - Label claims as supported, refuted or not enough info: [FEVER 2018](#) Shared Task; [FEVER 2.0 2019](#) Shared Task
The Language of Manipulation (Semi-structured)

- Identify and analyze polar opinions, framing and perspectives on social media or in partisan news corpora, e.g. Demszky et al. 2019 or Chen et al. 2019