Fairness in NLP

Dr. Rachael Tatman
Senior Developer Advocate @ Rasa

Strata Data & AI Superstream Series: Natural Language Processing
October 27, 2020

How I got here

- BA in Linguistics & English literature (William and Mary)
- PhD in Linguistics (University of Washington)
    - Phonetics & ASR
    - Computational Sociolinguistics
    - Ethics/FAT in NLP
- Data scientist/Developer advocate at Google (Kaggle)
- Developer advocate at Rasa
    - Open source Conversational AI framework

What’s most urgent in NLP?

- Where I started:
    - Language technologies should work equally well for everyone
- Where I am now:
    - Language technologists must proactively mitigate harm from our work

There are many different ways to think about fairness.

A quick intro to fairness in machine learning

- Lots of different definitions/approaches to fairness!
- Most critical when…
  - working with human data
  - in high stakes situations
- Legally necessary when dealing with (Barocas & Hardt 2017):
  - Credit (Equal Credit Opportunity Act)
  - Education (Civil Rights Act of 1964; Education Amendments of 1972)
  - Employment (Civil Rights Act of 1964)
  - Housing (Fair Housing Act)
  - ‘Public Accommodation’ (Civil Rights Act of 1964)

Some current frameworks:

- Counterfactual Fairness
- Equality of Opportunity
- Fairness Through Awareness
- Disparate Impact

Counterfactual Fairness

"A decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group" (Kusner et al. 2017)

Counterfactual Fairness

- Definition
  - “A decision is fair towards an individual if it the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group” (Kusner et al. 2017)
  - In other words, a person should have the same probability of being approved for a loan even if we hypothetically changed their race
- Criticisms
  - Can be brittle/Very sensitive to different models
  - Social group membership is not independently distributed

Equality of Opportunity

“Individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome” (Hardt et al. 2016)

Equality of Opportunity

- Definition:
    - “Individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome” (Hardt et al. 2016)
    - In other words, everyone who pays their bills on time should have the same probability of getting a loan, regardless of their race
- Criticisms
    - Based on the philosophical idea of equality of opportunity, which faces its own criticisms
    - Bias can arise in determining what counts as qualification

Fairness Through Awareness

“[A]ny two individuals who are similar with respect to a particular task should be classified similarly” (Dwork et al, 2011)

Fairness Through Awareness

- Definition
  - “[A]ny two individuals who are similar with respect to a particular task should be classified similarly” (Dwork et al, 2011)
  - An alternative to “fair affirmative action”/group fairness, which measures fairness as statistical parity with population-level measures
- Criticisms
  - Bias can be introduced through quantifying similarity
  - May lead to group-level unfairness
- How is this different from equality of opportunity?
  - Equality of opportunity only considers individuals above a certain floor (“qualification”), not a pairwise comparison of matched individuals

Disparate Impact

"[A] selection process has widely different outcomes for different groups, even as it appears to be neutral" (Feldman et al 2014)

Disparate Impact

- Definition
  - "[A] selection process has widely different outcomes for different groups, even as it appears to be neutral" (Feldman et al 2014), based on the 1964 Civil Rights act
  - Disparate impact is unintentional, disparate treatment is intentional discrimination
- Criticisms
  - May appear unfair to individuals (see Dwork et al, 2011)
  - Formally only applies to protected classes vs. those not in protected classes, may miss intersections

Some current frameworks:

- Counterfactual Fairness
  - The output for an individual should be the same as if they belonged to a different demographic groups
- Equality of Opportunity
  - All qualified individuals should have the same chance of success
- Fairness Through Awareness
  - Any two similar individuals should have the same outcome
- Disparate Impact
  - We should minimize unfair differences between groups that arise through neutral-seeming individual choices

Should a truly fair system have all of these qualities? Can we have such systems?

A fairness gap to fill

- These measures are applied at the level of the individual
- What about systems that perform some task for humans?
  - Speech recognition
  - Motion detection (e.g. for automatic faucets)
  - Facial recognition

My working approach:

- Parity in utility
  - **Systems should not perform tasks significantly more poorly for some demographic groups than others**
  - This is a goal & may pose significant technical challenges (lack of training data, handling physiological differences for tasks like ASR, etc.)

What Social factors matter? (Barocas & Hardt 2017)

Race (Civil Rights Act of 1964)
Color (Civil Rights Act of 1964)
Sex (Equal Pay Act of 1963, Civil Rights Act of 1964)
Religion (Civil Rights Act of 1964)
National origin (Civil Rights Act of 1964)
Citizenship (Immigration Reform and Control Act)
Age (Age Discrimination in Employment Act of 1967)
Pregnancy (Pregnancy Discrimination Act)
Familial status (Civil Rights Act of 1968)
Disability status (Rehabilitation Act of 1973, Americans with Disabilities Act of 1990)
Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act)
Genetic information (Genetic Information Nondiscrimination Act)

"Can I minimize differences in accuracy between subgroups" is less important than "should this be built at all"

@rctatman

No system is inevitable.
It must be built.
It is up to us to decide what
our labor will support.

2016

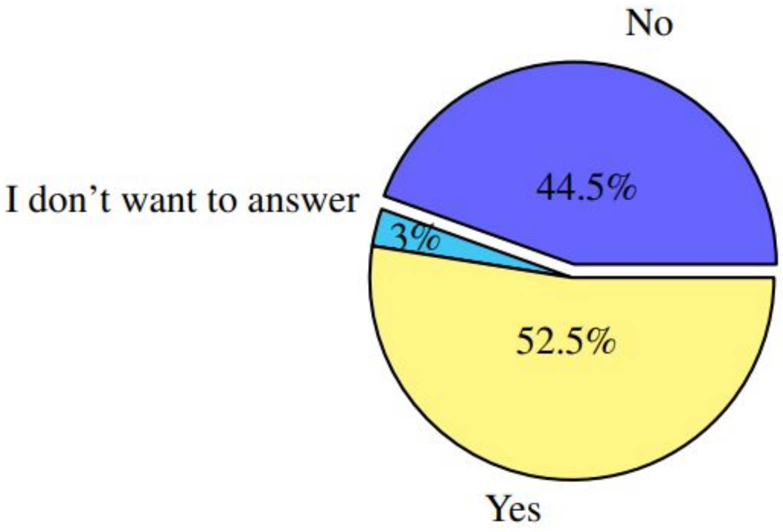


Figure 7: "Do you consider yourself responsible for the usages imagined from the applications/algorithms you create?" (international survey).

Fort, K., & Couillault, A. (2016, May). Yes, we care! results of the ethics and natural language processing surveys.

@rctatman

2016

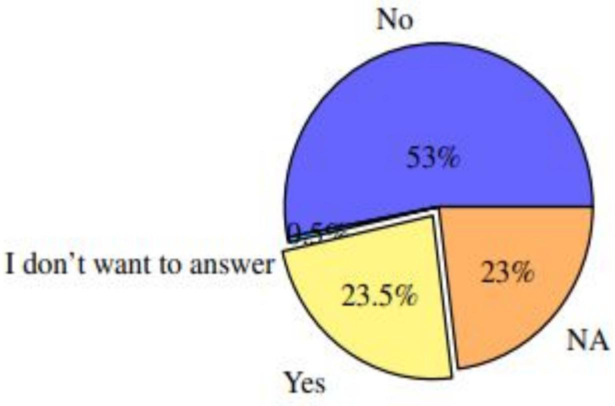


Figure 1: "Have you ever refused a project due to ethical issues?" (international survey).

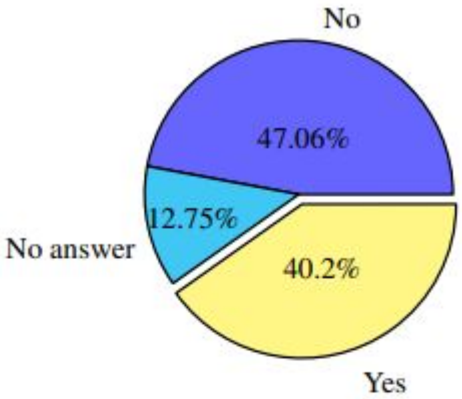


Figure 2: "Avez-vous déjà refusé ou limité un projet pour des raisons éthiques ? / Have you ever refused a project due to ethical issues?" (French survey).

Fort, K., & Couillault, A. (2016, May). Yes, we care! results of the ethics and natural language processing surveys.

2020: The ACL Adopted the ACM Code of Ethics

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing
- Avoid harm
- Be honest and trustworthy
- Be fair and take action not to discriminate
- Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.
- Respect privacy
- Honor confidentiality

2020: The ACL Adopted the ACM Code of Ethics

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing
- **Avoid harm**
- Be honest and trustworthy
- **Be fair and take action not to discriminate**
- Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.
- **Respect privacy**
- Honor confidentiality

What won't I help build? (My list)

- Surveillance technology
  - "**Peregrine E-mail Analysis Tool**: ingests bulk e-mail information and uses natural language processing (NLP) to help special agents and analysts extract keywords and run analytics", 2021 U.S. Immigration and Customs Enforcement Budget Overview
- Deceptive technology
  - Yang, Z., & Xu, C. (2019, November). *Read, Attend and Comment: A Deep Architecture for **Automatic News Comment Generation***. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP-IJCNLP**) (pp. 5080-5092).
  - All Rasa assistants identify themselves as bots
- Social category detectors

Social category detectors?

Systems that attempt to infer the social category of a user without the user voluntarily providing that information.

- Preoţiuc-Pietro, D., & Ungar, L. (2018, August). **User-level race and ethnicity predictors from twitter text**. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1534-1545).
- Ardehaly, E. M., & Culotta, A. (2015). **Inferring latent attributes of Twitter users with label regularization.** In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 185-195).
- Volkova, S., Coppersmith, G., & Van Durme, B. (2014, June). **Inferring user political preferences from streaming communications.** In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 186-196).
- Tatman, R., Stewart, L., Paullada, A., & Spiro, E. (2017, August). **Non-lexical features encode political affiliation on twitter.** In Proceedings of the Second Workshop on NLP and Computational Social Science (pp. 63-67).

Social category detectors?

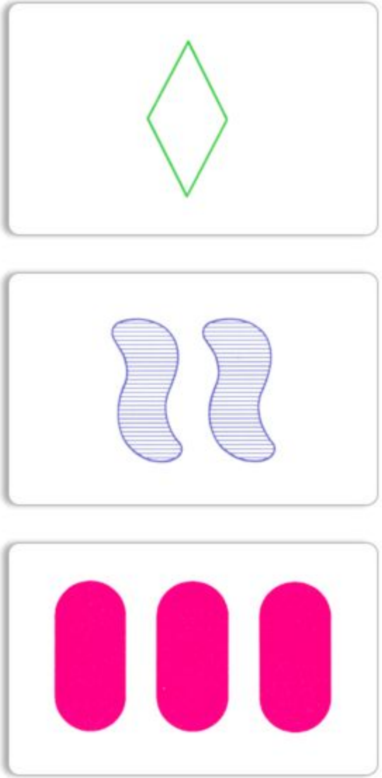
Systems that attempt to infer the social category of a user without the user voluntarily providing that information.

- Preoţiuc-Pietro, D., & Ungar, L. (2018, August). **User-level race and ethnicity predictors from twitter text**. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1534-1545).
- Ardehaly, E. M., & Culotta, A. (2015). **Inferring latent attributes of Twitter users with label regularization.** In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 185-195).
- Volkova, S., Coppersmith, G., & Van Durme, B. (2014, June). **Inferring user political preferences from streaming communications.** In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 186-196).
- Tatman, R., S & Spiro, E. (2017, August). **Non-lexical features encode political affiliation on twitter.** In Proceedings of the Second Workshop on NLP and Computational Social Science (pp. 63-67).

Yep, that's me.

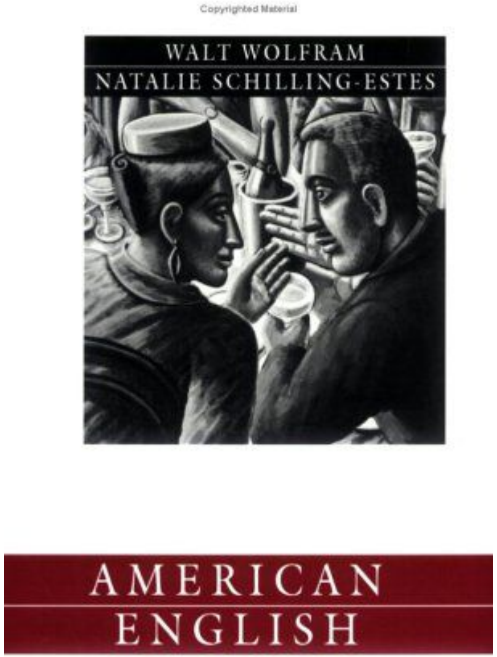
How to accidentally build a social category detector

1. Build an application that uses language data as input and assigns individuals into some sort of category
2. Don’t consider sociolinguistic variation
3. Don’t conduct socially-stratified validation
4. Voila!



How to accidentally build a social category detector

1. All language use is shaped by its social context
2. Many demographic factors are linked to systematic variation in language including:
   a. Regional Origin
   b. Age
   c. Socio-economic status/Social class
   d. Race/ethnicity



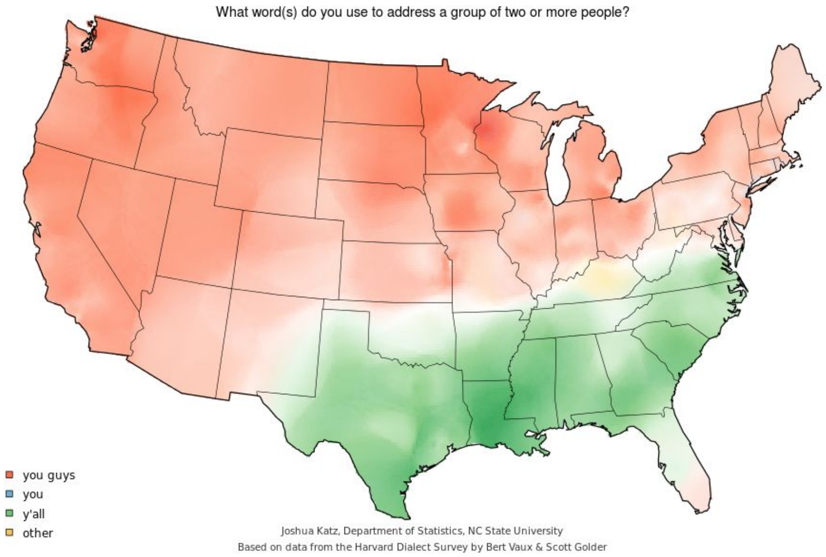
"American English" by Wolfram and Schilling-Estes is a nice introduction

@rctatman

How to accidentally build a social category detector

Sociolinguistic variation exists at every level of the grammar

a. Phonetics & sub-lexical features:
   i. cot/caught merger
   ii. :) vs. :-)
b. Lexical variation
   i. firefly vs. lightning bug
c. Syntactic variation
   i. Needs washed
   ii. We stay home anymore
d. Semantic variation
   i. lift, torch, boot


What word(s) do you use to address a group of two or more people?
🟥 you guys
🟦 you
🟩 y'all
🟨 other

Joshua Katz, Department of Statistics, NC State University
Based on data from the Harvard Dialect Survey by Bert Vaux & Scott Golder

@rctatman

How to NOT accidentally build a social category detector

- Extracted features > raw text input
- Consider relevant sociolinguistic variation
  - Quickest way = ask a sociolinguist familiar with your user population!
- Do sub-group validation for sociolinguistically-relevant groups
  - Which groups matter? Depends on who your users are.

Questions to Ask

1. Who benefits from this system existing?
2. Who could be harmed by this system?
3. Can users choose not to interact with this system?
4. Does that system enforce or worsen systemic inequalities?
5. Is this genuinely bettering the world? Is it the best use of your limited time and resources?

Supporting each other

- Decide ahead of time what your personal priorities are
- As researchers and practitioners we can affect change much faster than any legislature
- Stay informed and educate others about the risks and drawbacks of NLP applications
- Support each other in speaking up, coordinated effort gets results!
    - “Google reportedly leaving Project Maven military AI program after 2019” due in part to coordinated pressure from employees
    - Google’s choice not to offer facial recognition for surveillance was due in part to employee activism

Further Reading

- "The social impact of natural language processing" Hovy & Spruit, ACL 2016
- "Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis?" Leins, Lau & Baldwin, ACL 2020
- "Ethics and Data Science" Loukides, Mason & Patil, 2018
- "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms" FAT/ML