Natural Language Processing

Lecture 21: Ethics & Safety for NLP

Many slides & instruction ideas borrowed from:
Mohit Iyyer, David Bamman & Greg Durret

Expanding Scope of NLP Models

- The impact of NLP is growing universally
    - People interact with NLP systems extensively
        - Conversational agents
    - People make decisions based on NLP systems
        - Ads, recommender systems, …
    - Legal decisions are being impacted by NLP systems.
        - Paroles, immigration, etc.

Expanding Scope of NLP Models

In 2020, Uma Mirkhail got a firsthand demonstration of how damaging a bad translation can be.

A crisis translator specializing in Afghan languages, Mirkhail was working with a Pashto-speaking refugee who had fled Afghanistan. A U.S. court had denied the refugee's asylum bid because her written application didn't match the story told in the initial interviews.

In the interviews, the refugee had first maintained that she'd made it through one particular event alone, but the written statement seemed to reference other people with her at the time — a discrepancy large enough for a judge to reject her asylum claim.

After Mirkhail went over the documents, she saw what had gone wrong: An automated translation tool had swapped the "I" pronouns in the woman's statement to "we."

https://restofworld.org/2023/ai-translation-errors-afghan-refugees-asylum/

Expanding Scope of NLP Models

Expanding Scope of NLP Models

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

Alexa tells 10-year-old girl to touch live plug with penny

28 December 2021

Share

Ethics and Safety for NLP

- WWII —> Regulation in experimentations with human subjects
    - IRB
- For NLP/AI?
-

Ethics & Safety for NLP

- Amplifying the existing Bias
- Exclusion of the underprivileged
- Risks in automation
- Unethical use: harmful usage of systems

Hovy and Spruit ACL 2016

Ethics & Safety for NLP

- **Amplifying the existing Bias**
- Exclusion of the underprivileged
- Risks in automation
- Unethical use: harmful usage of systems

Hovy and Spruit ACL 2016

Word Embeddings

- Mikolov et al. 2013 show that vector representations have some potential for analogical reasoning through vector arithmetic.

$$\text{apple} - \text{apples} \approx \text{car} - \text{cars}$$

$$\text{king} - \text{man} + \text{woman} \approx \text{queen}$$

Mikolov et al., (2013), “Linguistic Regularities in Continuous Space Word Representations” (NAACL)

Science AAAS

Home News Journals Topics Careers

Science Science Advances Science Immunology Science Robotics Science Signaling Science Translational Medicine

REPORT

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]

+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

Peer Reviewed
see details

Bias

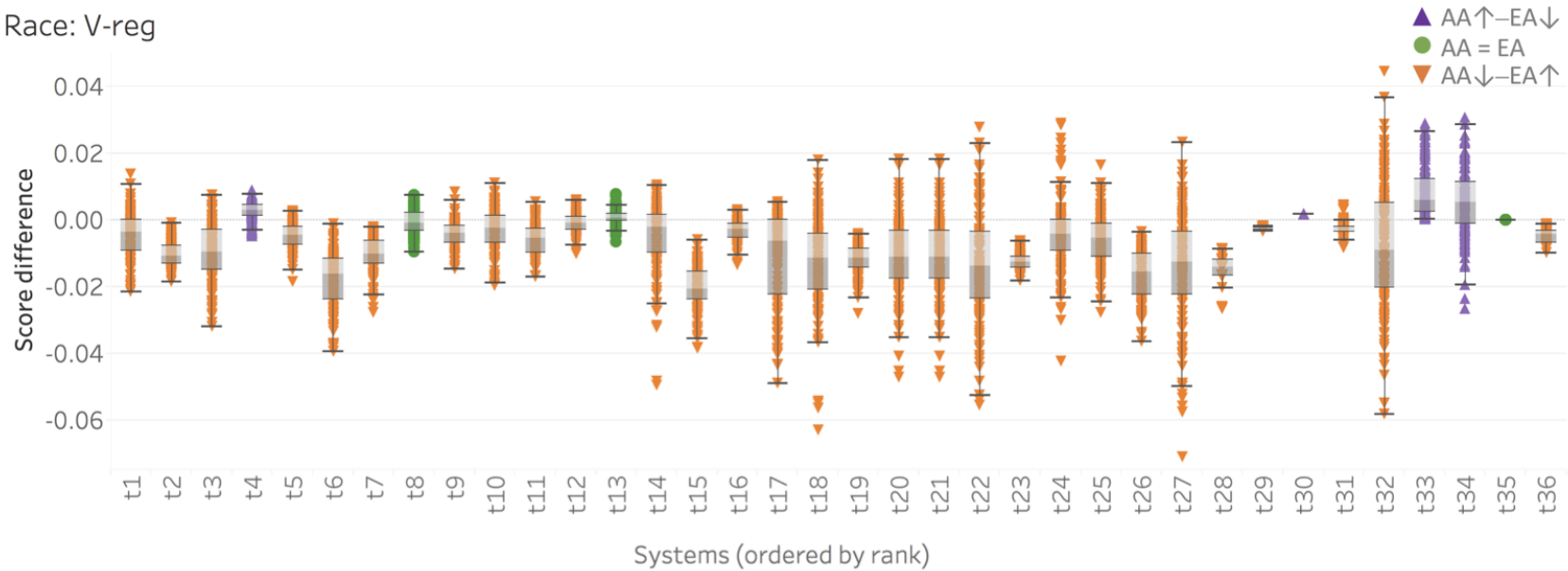
- Allocational harms: automated systems allocate resources unfairly to different groups (access to housing, credit, parole).
- Representational harms: automated systems represent one group less favorably than another (including demeaning them or erasing their existence).

Blodgett et al. (2020), “Language (Technology) is Power: A Critical Survey of “Bias” in NLP”

Representations

- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit.

- Embeddings for African-American first names are closer to “unpleasant” words than European names (Caliskan et al. 2017)



- Sentiment analysis over sentences containing African-American first names are more negative than identical sentences with European names

Kiritchenko and Mohammad (2018), "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems"

Amplifying the Bias: LLMs

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

- Bias in the data: Model “size does not guarantee diversity”
-

Amplifying the Bias: LLMs

- RLHF reduces to some extent.

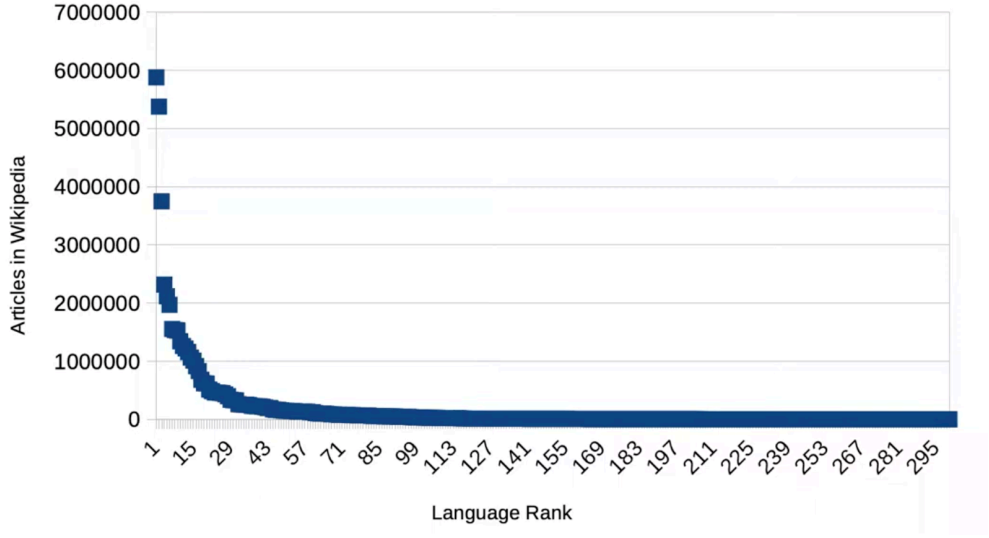
Ethics & Safety for NLP

- Amplifying the existing Bias
- **Exclusion of the underprivileged**
- Risks in automation
- Unethical use: harmful usage of systems

Hovy and Spruit ACL 2016

Exclusion of the underprivileged

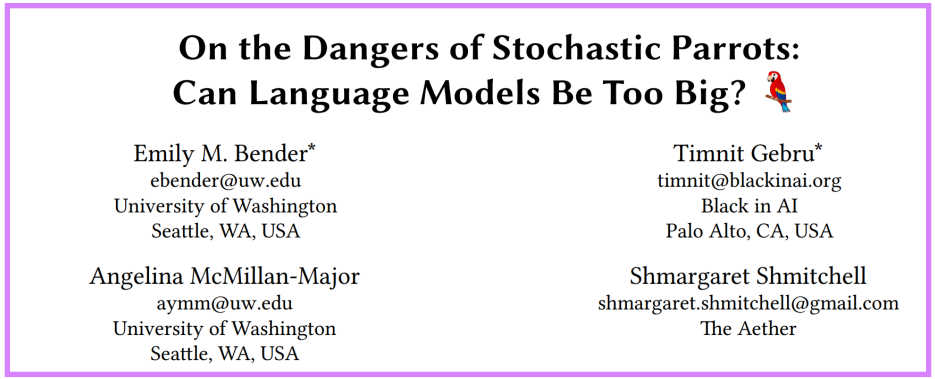
- NLP (and internet content) is predominantly is English-centric
-



Exclusion of the underprivileged

- Presence of people outside the main-stream:
    - Dialects
    - Minorities
    - Elderly
- Treatment of data annotators & turkers
- The burden of cost

Exclusion of the underprivileged



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

- Environmental cost: disproportionately on marginalized populations, who aren't even well-served by models.
- Massive data is challenging to audit, contains data that is biased and is mostly a snapshot of a single point in time.

Ethics & Safety for NLP

- Amplifying the existing Bias
- Exclusion of the underprivileged
- **Risks in automation**
- Unethical use: harmful usage of systems

Hovy and Spruit ACL 2016

Risks in Automation

- NLP systems are impacting decision making


THE VERGE TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE

US & WORLD TECH POLITICS

Facebook apologizes after wrong translation sees Palestinian man arrested for posting 'good morning'

Facebook translated his post as 'attack them' and 'hurt them'

by Thuy Ong | @ThuyOng | Oct 24, 2017, 10:43am EDT

In 2020, Uma Mirkhail got a firsthand demonstration of how damaging a bad translation can be.

A crisis translator specializing in Afghan languages, Mirkhail was working with a Pashto-speaking refugee who had fled Afghanistan. A U.S. court had denied the refugee's asylum bid because her written application didn't match the story told in the initial interviews.

In the interviews, the refugee had first maintained that she'd made it through one particular event alone, but the written statement seemed to reference other people with her at the time — a discrepancy large enough for a judge to reject her asylum claim.

After Mirkhail went over the documents, she saw what had gone wrong: An automated translation tool had swapped the "I" pronouns in the woman's statement to "we."

- Risks of LLM-based annotation
- Risks of LLM-generated WWW.

Risks in Automation

- The impact on the labor market
-

Risks in Automation

- The impact on the labor market
-

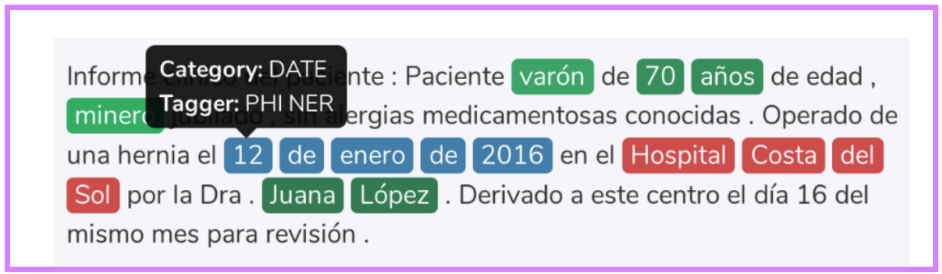
Ethics & Safety for NLP

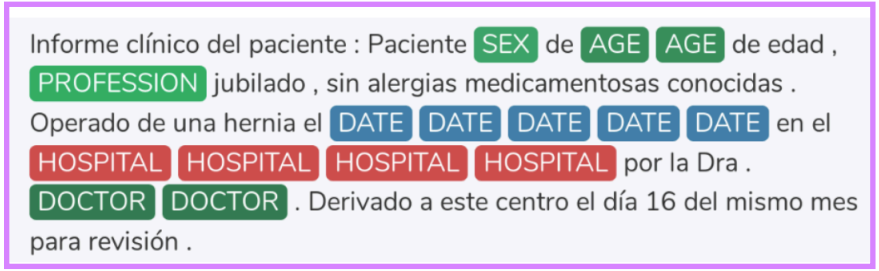
- Amplifying the existing Bias
- Exclusion of the underprivileged
- Risks in automation
- **Unethical use: harmful usage of systems**

Hovy and Spruit ACL 2016

Unethical use of NLP

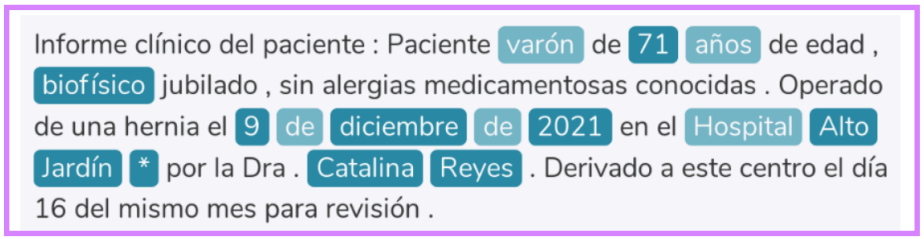
- Surveillance Systems
- Authorship attribution and de-anonymization

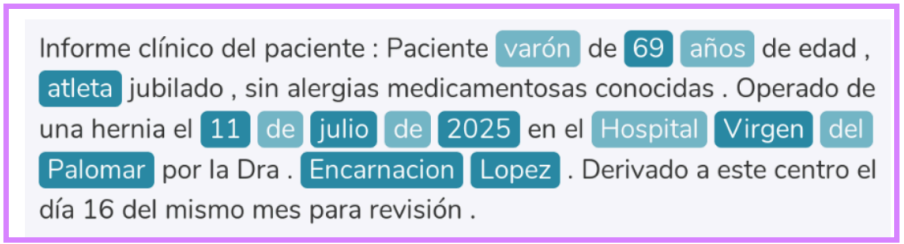




Unethical use of NLP

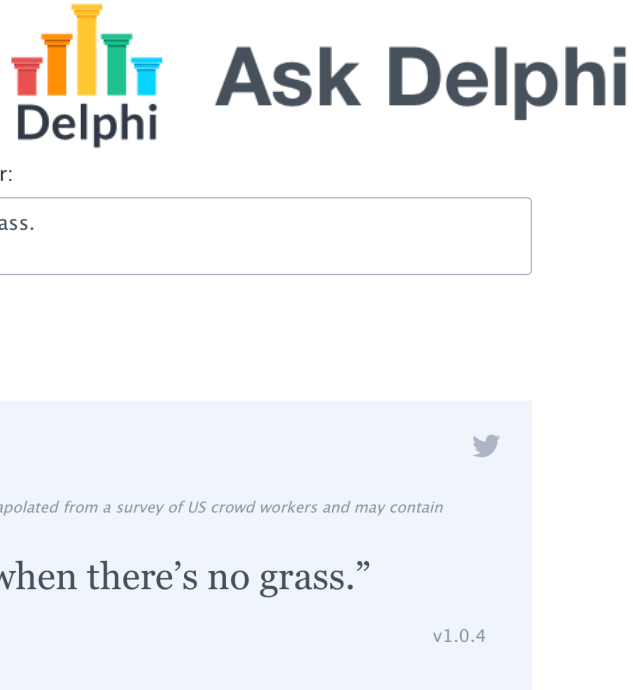
- Surveillance Systems
- Authorship attribution and de-anonymization





Ethics in NLP

- AI2's Delphi:



https://delphi.allenai.org/

Ethics & Safety for NLP

- Amplifying the existing Bias
- Exclusion of the underprivileged
- Risks in automation
- Unethical use: harmful usage of systems

Hovy and Spruit ACL 2016

Moving Forward