Detecting and Rewriting Socially Biased Language

Maarten Sap he/him

@ PAN '21

Previously

UWNLP

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Currently

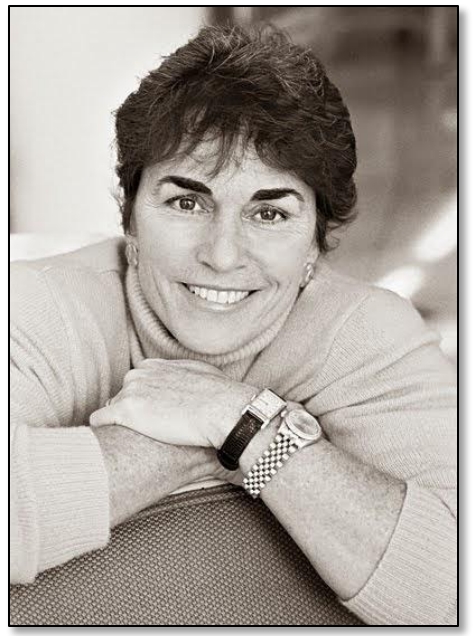
AI2 MOSAIC COMMONSENSE

Fall 2022

Carnegie Mellon University Language Technologies Institute

Language reflects society

Language is the road map of a culture.
It tells you where its people come from and where they are going.



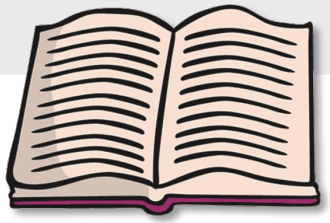
Rita Mae Brown
American feminist writer
& LGBTQ+ activist

Cycle of social inequality in text

World influences language

Text

- Language reflects existing social inequality [Lakoff, '73; Fiske, '93]
- Portrayals of minority characters are biased and stereotypical
  [e.g., Fast et al. '16; Field & Tsvetkov '19]
- Hate speech and toxicity mostly targets minority groups
  [e.g., Cleland, 2014; Trindade, 2018]



World

- Social inequality between majority-minority groups
  [e.g., RWJ Foundation, '17]
- Hate speech worsens relationships between groups [Bilewicz & Soral, '20]
- Portrayal of minority characters shapes perception and stereotypes
  [Behm-Morawitz & Mastro, '08; Field et al. '19]

Language influences world

In this talk:

How can we make natural language processing (NLP) systems **understand and mitigate social biases and toxicity in language**

Why this is important:

Any human-generated data **reflects social dynamics/inequality**, and NLP systems that **fail to account** for that **risk harmful results**

Social bias in Natural Language Processing (NLP)

Conversational AI

digital assistants, chatbots

Language generation

text autocompletion, story continuation

Text understanding

hate speech detection, sentiment analysis

Failure modes

toxic, rude, or **offensive** conversations


THE VERGE TECH SCIENCE ENTERTAINMENT MORE

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

mindless, **biased**, or **offensive** text generation


IEEE SPECTRUM

OpenAI's GPT-3 Speaks! (Kindly Disregard Toxic Language)

By Eliza Strickland

Posted 01 Feb 2021 | 17:03 GMT

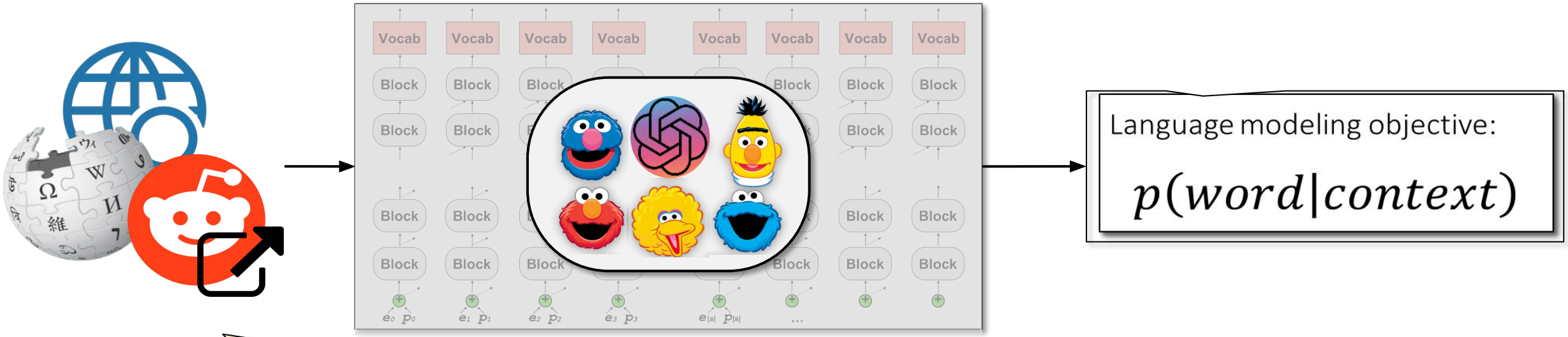
worse performance on minority user input, **biased** behavior


Forbes

Aug 13, 2019, 11:45am EDT | 11,211 views

Google's Artificial Intelligence Hate Speech Detector Is 'Racially Biased,' Study Finds

Pretrained language models (PTLMs)



Pretraining corpora:

- Documents from English Wikipedia, books (GPT, BERT)
- Documents from **outbound links from Reddit** (GPT-2)
- Archives of **all documents on the internet** (T5, GPT-3)

Learning language from random internet data…

what could go wrong? 

Mindless and socially-oblivious PTLMs

Pretrained LMs are…

- **learning stereotypes and social biases** from training data
  [Lakoff, '73; Fiske, '93; Fast et al. '16; Sheng et al. '19; Nadeem et al. '20; Bender et al. '21]
- at **risk for generating toxicity** in <100 generations
  [Gehman, Gururangan, ***Sap***, et al. '20]

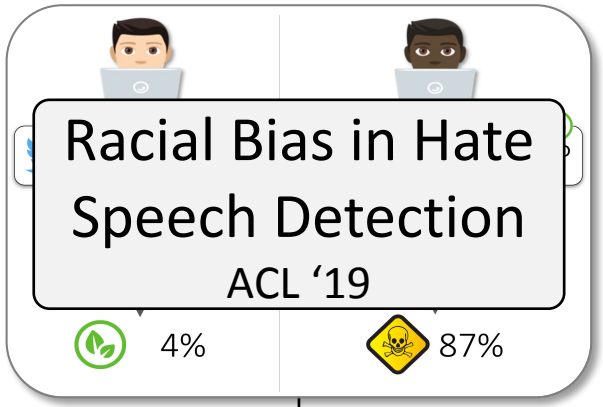
"Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy"

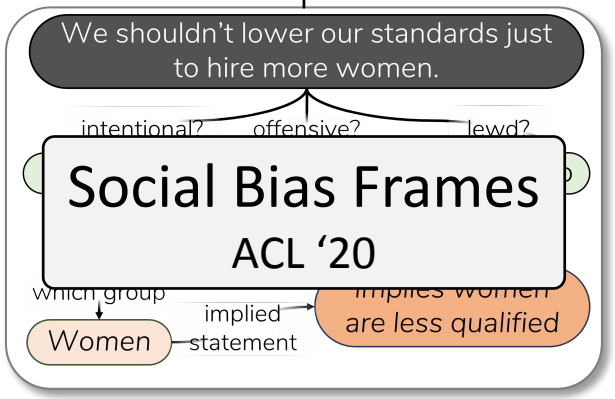
Prof. Ruha Benjamin, PhD

We need **formalisms to represent and detect social biases** and **algorithms to mitigate and avoid biased implications**

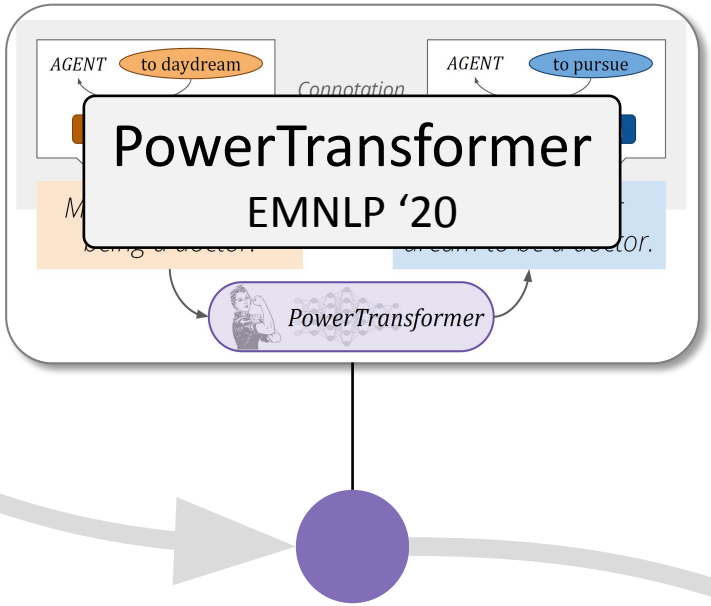
Talk outline

Detecting toxicity and social biases



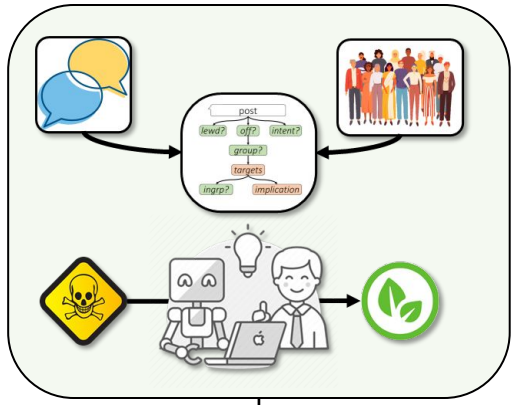


Rewriting and debiasing text



Future work

Human-centric social bias detection and mitigation

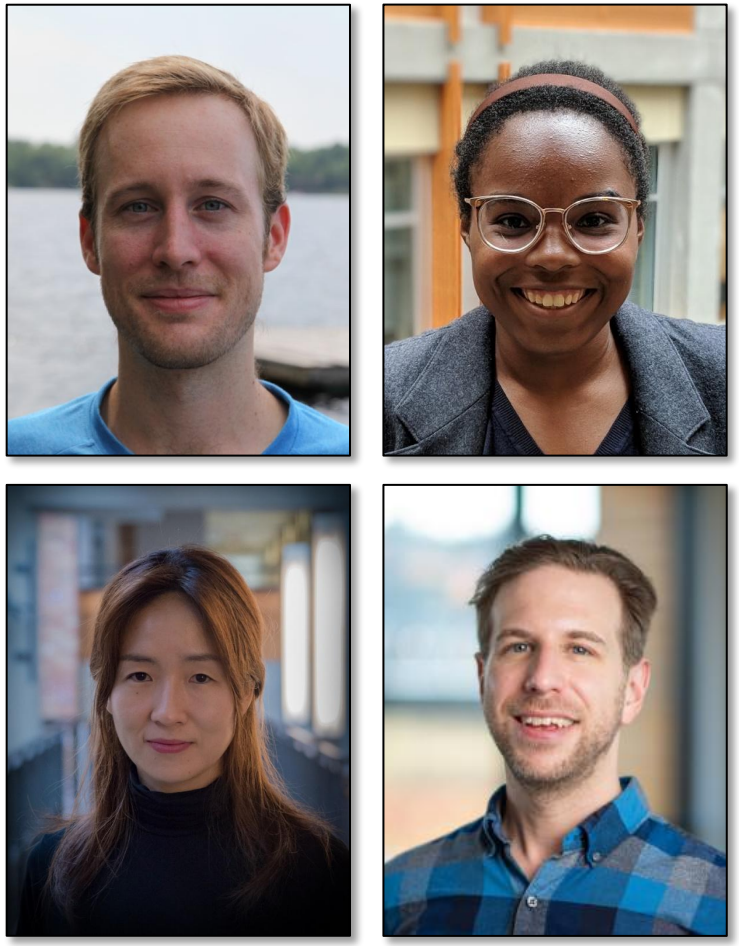


The Risk of Racial Bias in Hate Speech Detection

Maarten Sap, Dallas Card, Saadia Gabriel,
Yejin Choi & Noah A Smith

ACL 2019

☆ ***best paper nomination*** ☆



Hate speech is rampant online

billboard

POP

Fifth Harmony's Normani Quits Twitter Over Racist Bullying: 'I Can't Subject Myself Any Longer to the Hate'

8/8/2016 by Gil Kaufman

NATIONAL REVIEW

Jan. 14, 2020 SUBSCRIBE

Twitter Mob Attacks Transgender YouTube Star

By MADELEINE KEARNS | September 6, 2019 10:08 AM

BBC News Sport More

NEWS

Transgender people treated 'inhumanely' online

By Ben Hunte

BUSINESS INSIDER Subscribe

Amazon, Facebook, Twitter, and YouTube are all facing moderation issues — here's how America's tech giants are struggling to police their massive platforms

Rebecca Aydin Aug 24, 2019, 7:23 AM

USA TODAY

News Sports Entertainment Life Money [Tech] Travel Opinion

What civil rights groups want from Facebook boycott: Stop hate speech and harassment of Black users

Jessica Guynn USA TODAY

Published 3:01 a.m. ET Jul. 7, 2020 | Updated 6:07 p.m. ET Jul. 8, 2020

Platforms struggle to moderate content

Challenging to rely on humans solely

- **Scale** prevents effective centralized responses (e.g., Twitter, Facebook)
- Delegating to community (e.g., Reddit) can lead to **hate-endorsing communities**
- Content moderators suffer **inhumane** working conditions, **mental health issues**, etc.

Wulczyn et al. 2017

Kapoor et al. 2018

Ribeiro et al. 2018

Park et al. 2018

Waseem et al. 2018

van Aken et al. 2018

Davidson et al. 2017

Alorainy et al. 2018

Dixon et al. 2018

Zhang et al. 2018

Schmidt and Wiegand 2017

Golbeck et al. 2017

Park and Fung 2017

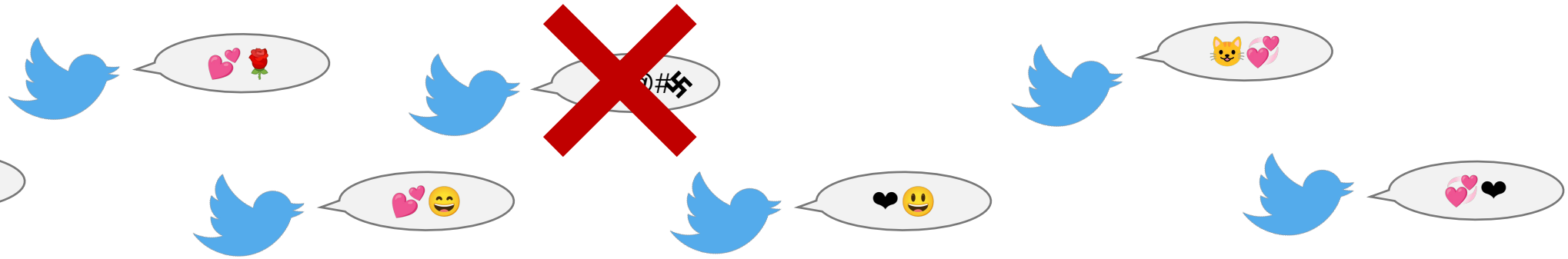
Founta et al. 2018

Kwok and Wang 2013

Burnap and Williams 2015

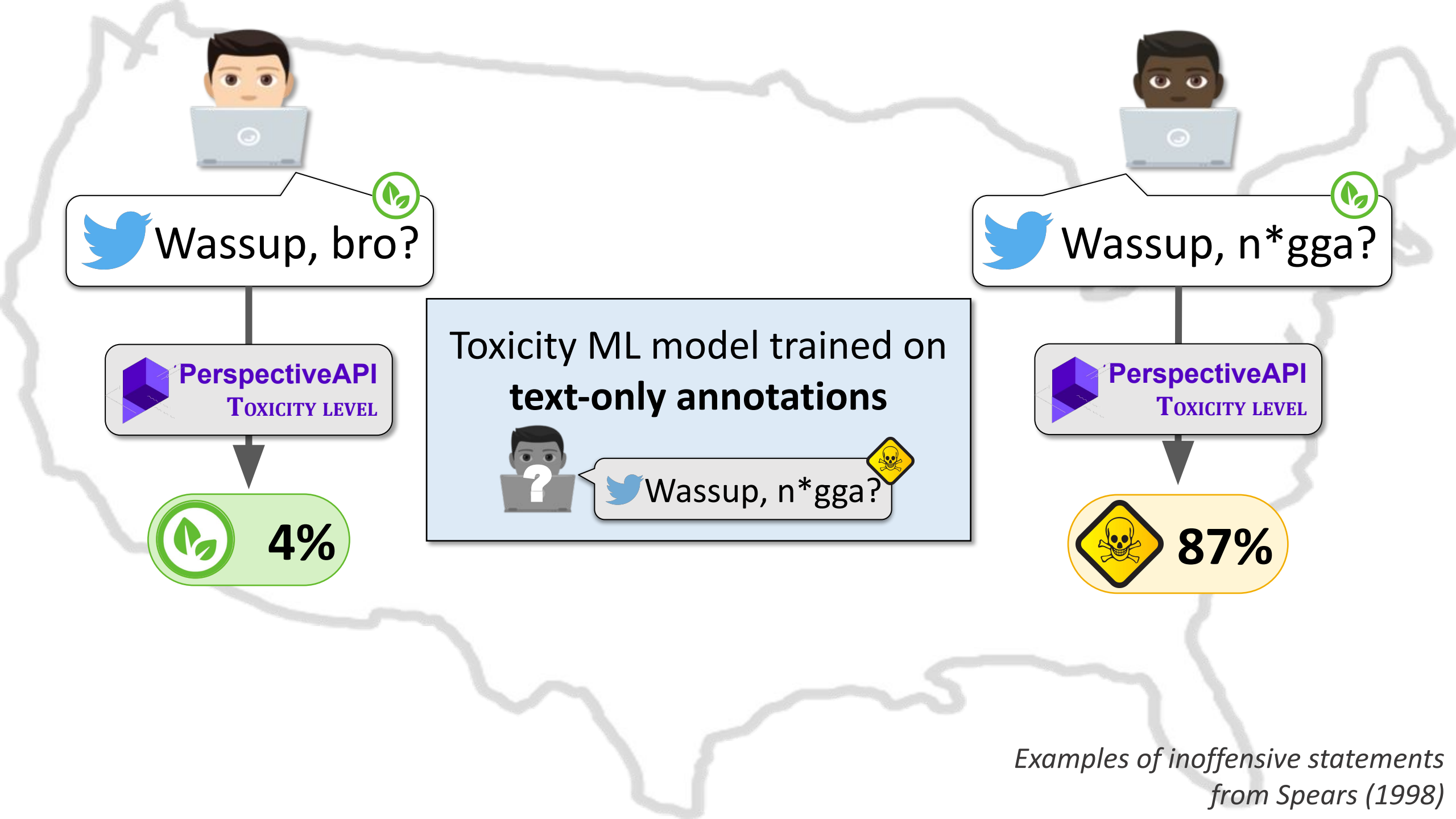
PerspectiveAPI
TOXICITY LEVEL

Automatic hate speech detection: try and make the internet **less toxic**



Problem: severe racial bias in hate speech detection

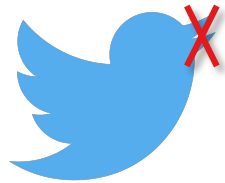
Hate speech against racial minorities?



Examples of inoffensive statements from Spears (1998)

Datasets **ignore underlying social dynamics** of speech
(e.g., identity of speaker, dialect of English)

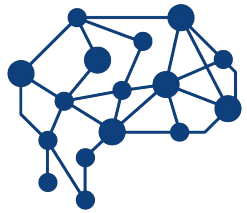
Ignoring theses nuances risks **harming minority populations** by suppressing inoffensive speech



We characterize and quantify **racial bias** in hate speech detection

Investigating racial bias in hate speech detection

1. Do **ML models acquire** this racial bias from datasets?



2. Can **annotation task design** affect these racial biases?

Why *racial* bias?

- Minority populations are most often the **target of hate speech** (Cleland, 2014; Trindade, 2018)
- **Racial bias less studied** than other identity-based biases
  - E.g., gender (Park et al., 2018; Gonen & Goldberg, 2019; Sun et al., 2019; Masahiro Kaneko & Bollegala, 2019; Zmigrod et al., 2019; Sweeny & Najafian, 2019; Stanovsky et al., 2019)
- On Twitter: **danger of silencing Black folks** disproportionately
  - Cultural importance, "*Black Twitter*" (Williams & Domoszlai, 2013)
  - Space for activism (Freelon et al., 2016; Anderson et al., 2018)

Challenge: Twitter profiles don't have race associated with them

Image source: *http://henryjenkins.org/*

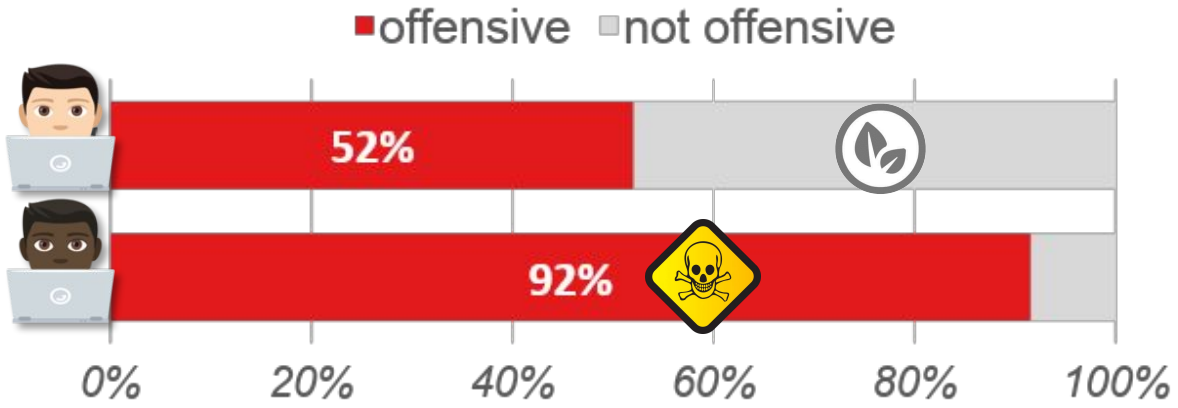
Dialect as proxy for racial identity

- *Premise*: specific lexical indicators of minority identity
- **African American English (AAE)** dialect
  - Common among (but not limited to) Black/African-American folks in U.S.
  - Variety of English that's extensively studied by linguists
  - Presence of AAE variants on Twitter [Jones, 2015]
- Lexical detector by Blodgett et al. (2016) to **infer presence of AAE**
  - *Note*: dialect/race much more complex

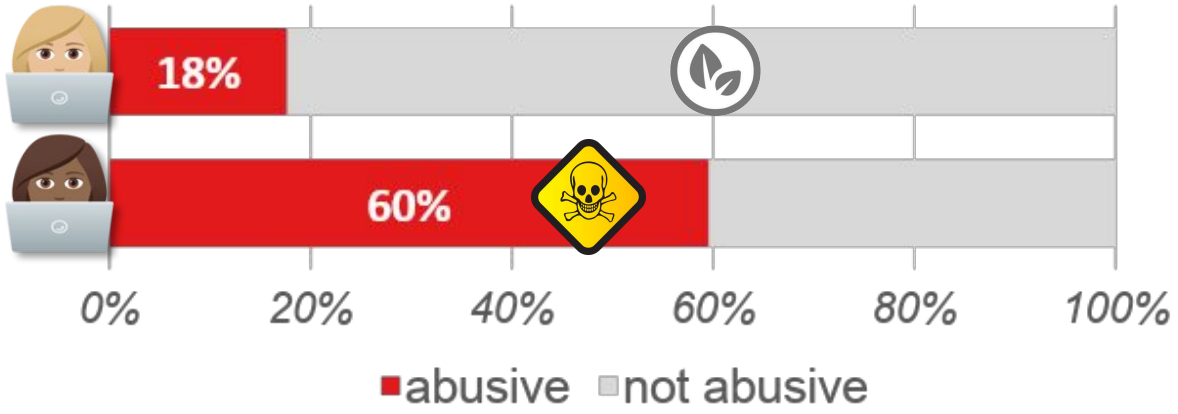
Image source: *https://www.languagejones.com/blog-1/2014/9/26/big-data-and-black-twitter*

How **racially biased** are datasets?

Proportions of toxic tweets in existing datasets



TWT-HATEBASE
(Davidson et al., 2017)



TWT-BOOTSTRAP
(Founta et al., 2018)

Bias also present in other datasets (see paper, or Davidson et al. 2019)

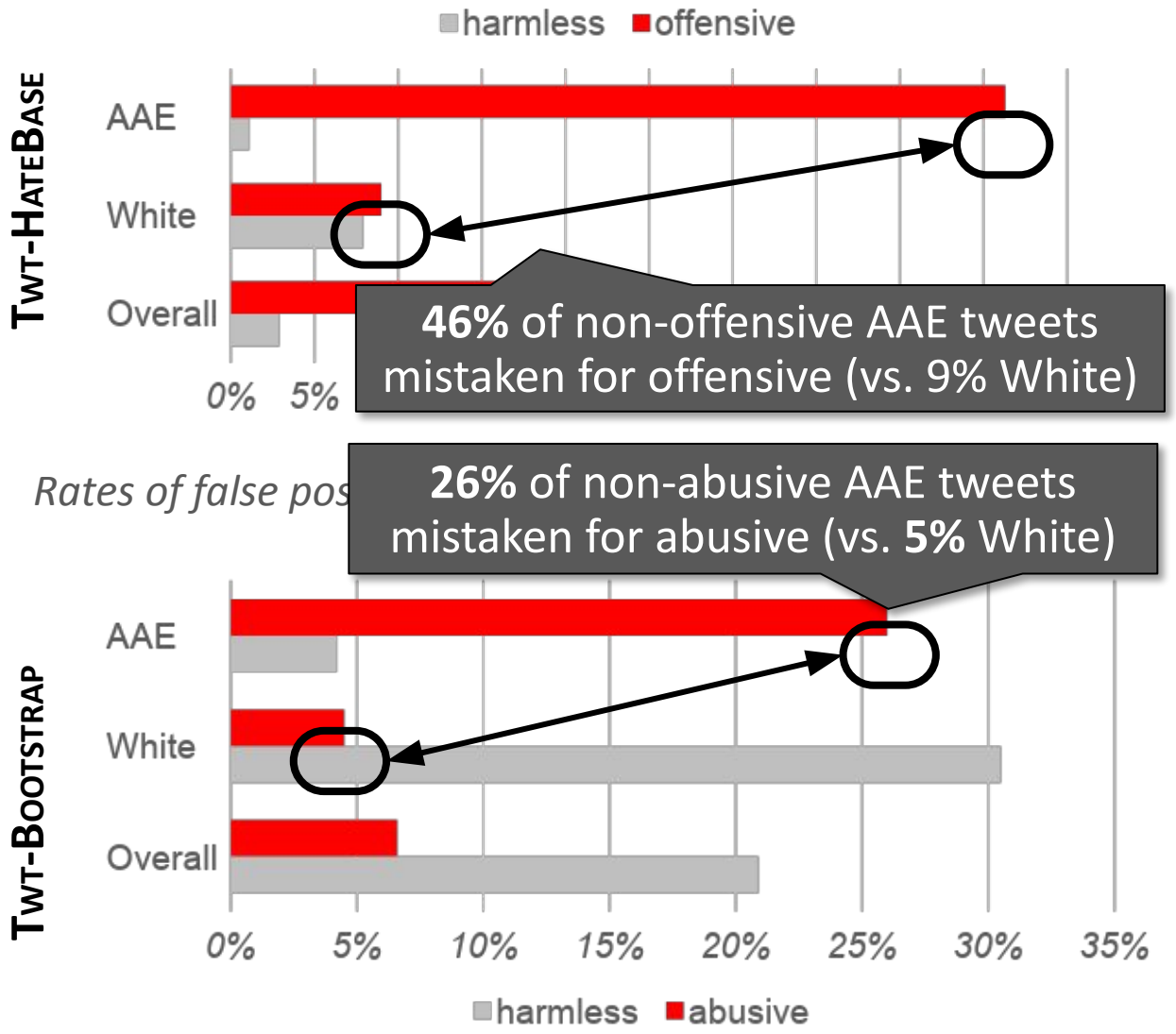
Q: Do ML models **acquire racial biases** from datasets?

A: They **acquire** and **exacerbate** them.

How are ML models affected by racial bias in datasets?

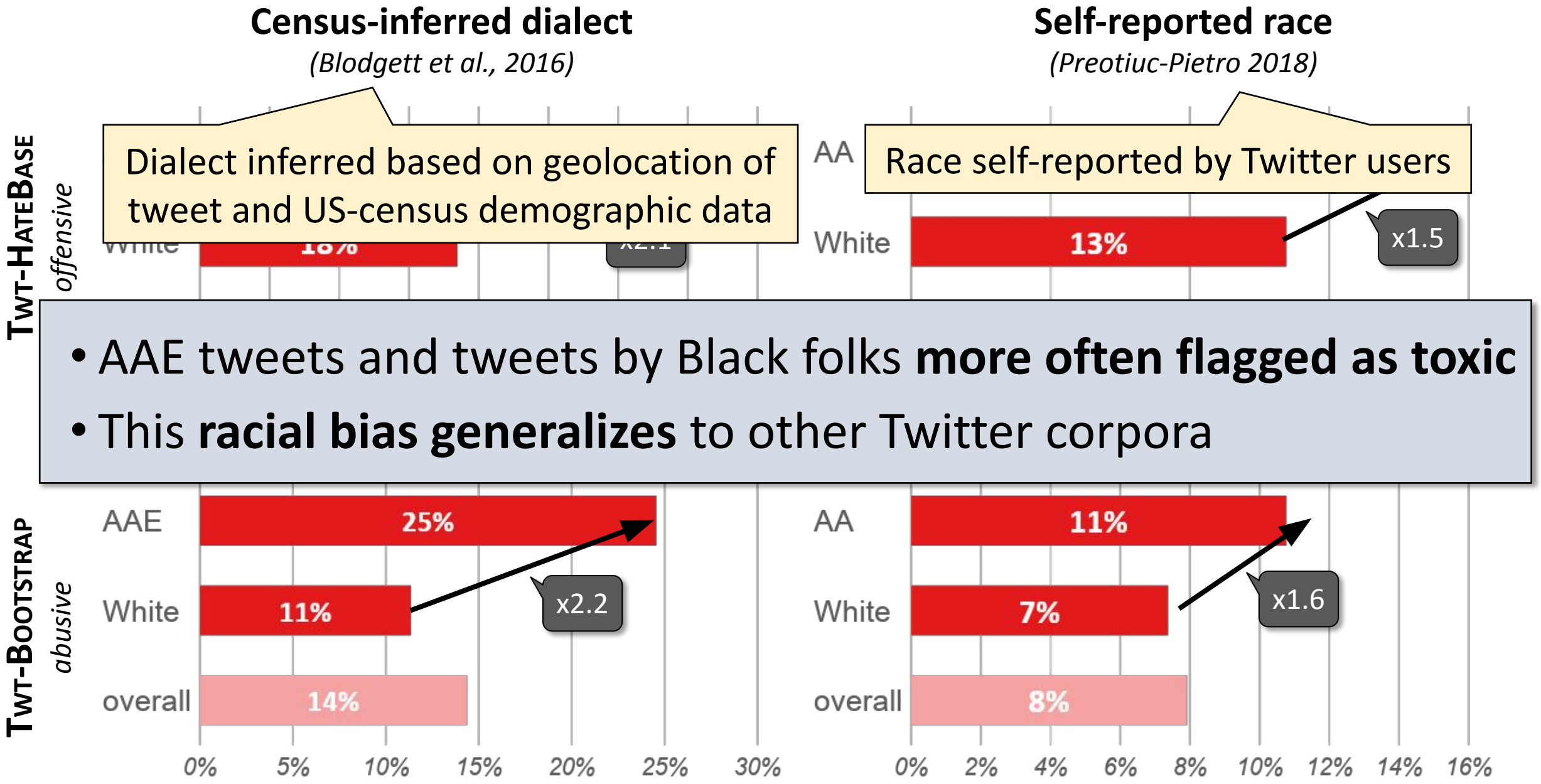
- Train/test two different classifiers
  - TWT-HATEBASE (Davidson et al, 2017)
  - TWT-BOOTSTRAP (Founta et al., 2018)
- Rates of **false flagging of toxicity**
  - Broken down by dialect group on heldout set
  - Equality of opportunity criterion [Hardt et al., 2016]

Predictions by both classifiers **biased against AAE tweets**



Q: Does this racial bias **generalize** beyond these datasets?

A: Unfortunately, it does…



Average classifier probability of offensive/abusive, broken down by dialect/group

Q: What can we do to reduce these biases?

A: Changing the data collection helps

Control condition

Text-only, no context

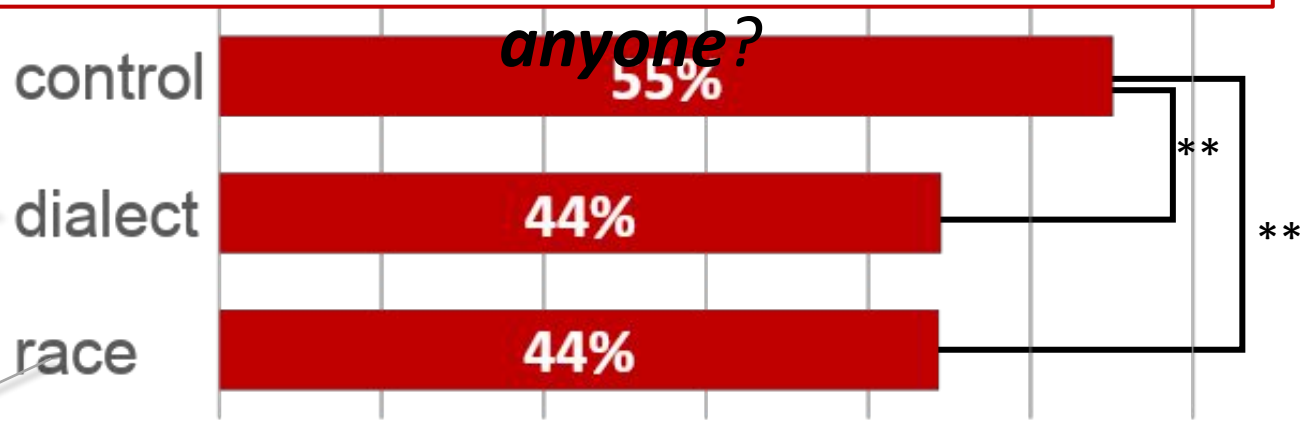
Dialect priming

"Our AI thinks this tweet is in African American English"

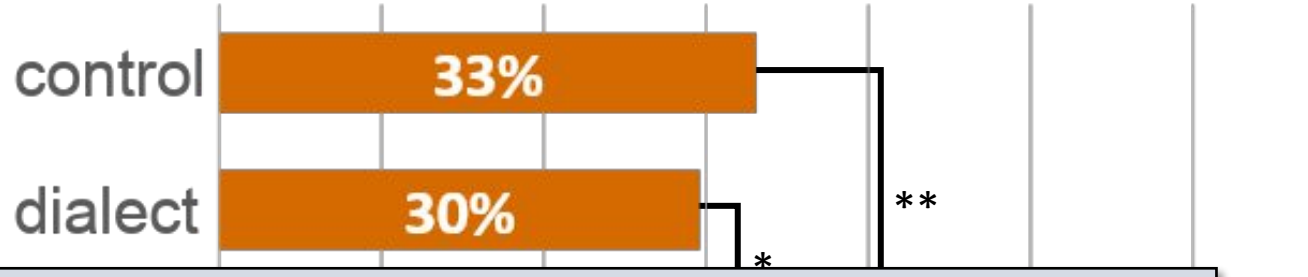
Race priming

"A Twitter user that is likely Black/African American tweeted…"

Could this tweet be **offensive to *anyone*?**



Is this tweet **offensive to *you*?**



MTurk study:

- [redacted]
- [redacted]
- [redacted]
- [redacted]

- Priming annotators about dialect/race **influences labels**
- Annotations of offensiveness are **highly subjective**

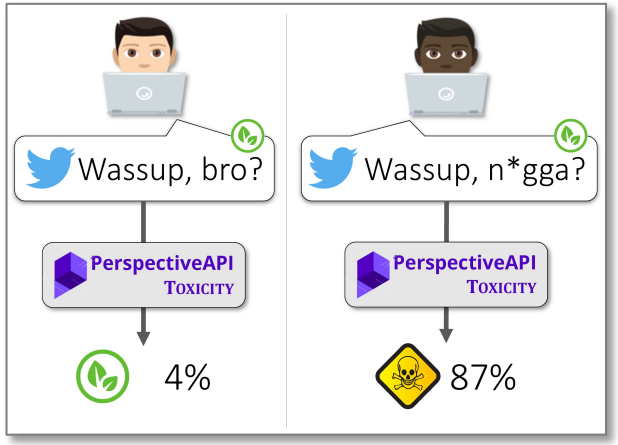
% <0.01

** p<0.001

Takeaways

Overt toxicity detection backfires against racial minorities

- Dialect-based **racial bias** in datasets, likely due to negative perception of race and AAE [Spears, '98; Rosa & Flores, '17]
- NLP models trained on biased data will **exacerbate the racial bias**
- Pilot study shows **highlighting dialect influences labels** of offensiveness



Dialect priming

"Our AI thinks this tweet is in African American English"

Race priming

"A Twitter user that is likely Black/African American tweeted…"

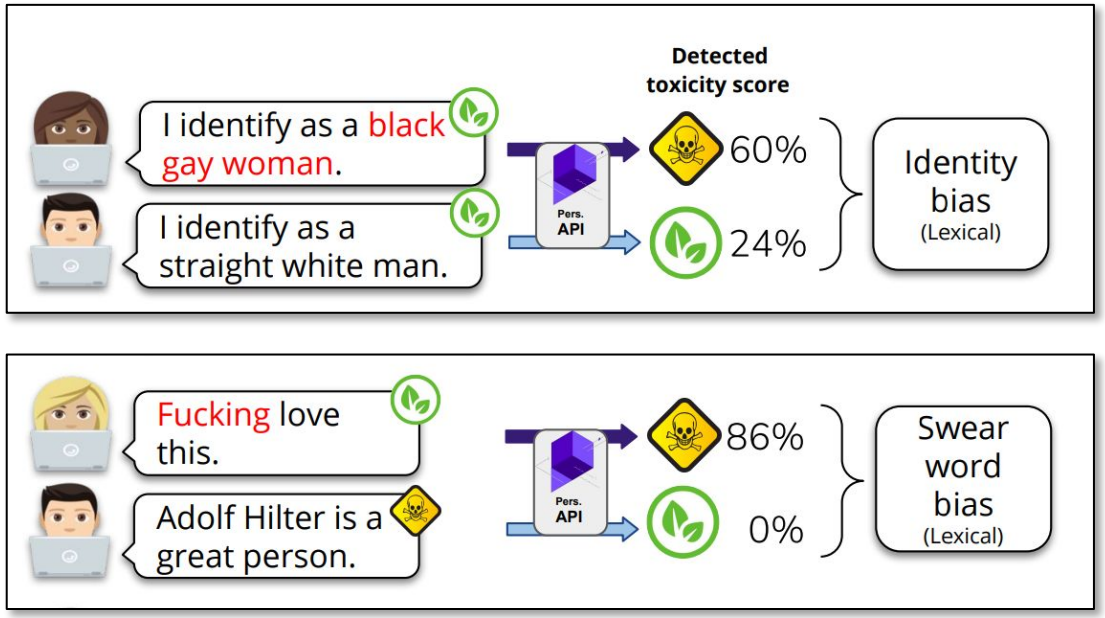
Maybe we should re-think how we tackle hate speech detection?

Biases in toxicity classification

- Besides racial bias, also lexical biases:
  - against minority identities
    [Park and Fung, 2017; Dixon et al., 2018]
  - against swearwords
    [Dinan et al. '19, Vidgen et al. '21]

In [Zhou, **Sap**, et al., EACL '21]:

- Can automatic debiasing methods from NLI mitigate these biases?
  - Ensemble learning, data filtering, etc.
- *Short answer*: easier if the bias is lexical (keywords) vs. **harder for dialect-based biases**



Challenges in Automated Debiasing for Toxic Language Detection

Xuhui Zhou[♡] Maarten Sap[♣] Swabha Swayamdipta[◇] Noah A. Smith[♣◇] Yejin Choi[♣◇]

[♡]Department of Linguistics, University of Washington
[♣]Paul G. Allen School of Computer Science & Engineering, University of Washington
[◇]Allen Institute for Artificial Intelligence
xuhuizh@uw.edu,{msap,yejin,nasmith}@cs.washington.edu, swabhas@allenai.org

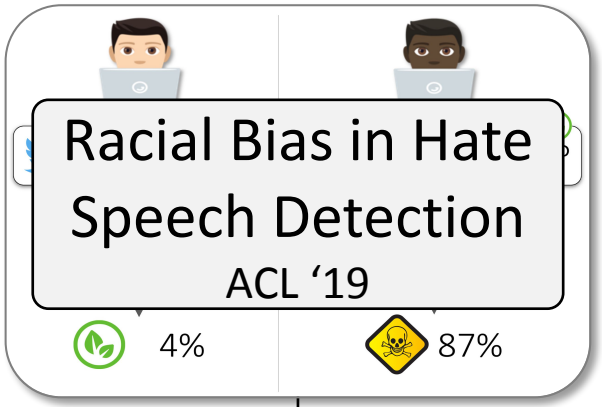
Who decides what is offensive or not?

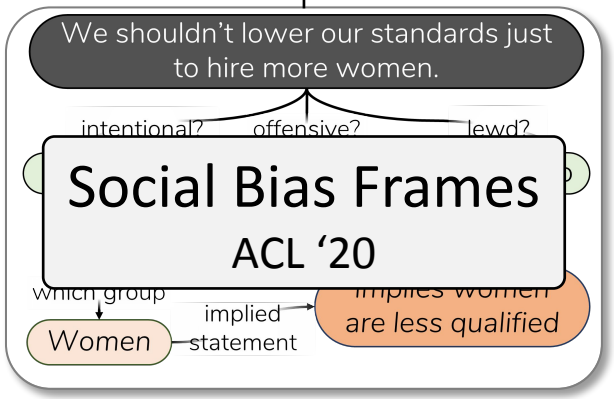
- **Subjectivity in annotations**
  - Annotator agreement is moderate
    [Ross et al., 2017, Waseem 2016]
  - Different perception based on your attitudes
    [Cowan & Khatchadourian, 2003; Kocon et al., 21]
  - Different based on identity
    [Breitfeller et al., 19; Larimore et al., 21]
- **Different definitions**: NLP vs. legal definitions
- **Missing explanations**: why is something hate or offensive speech?

*Binary classification is **not expressive enough** to capture and explain the full range of offensiveness in text*

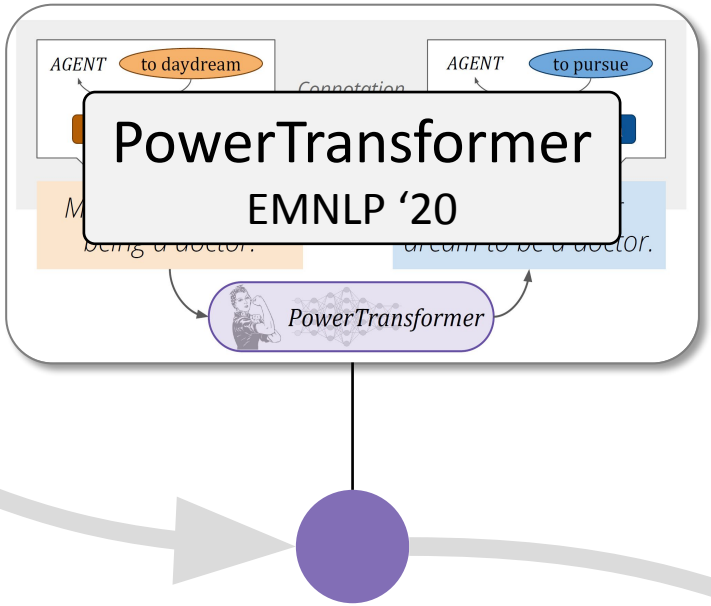
Talk outline

Detecting toxicity and social biases



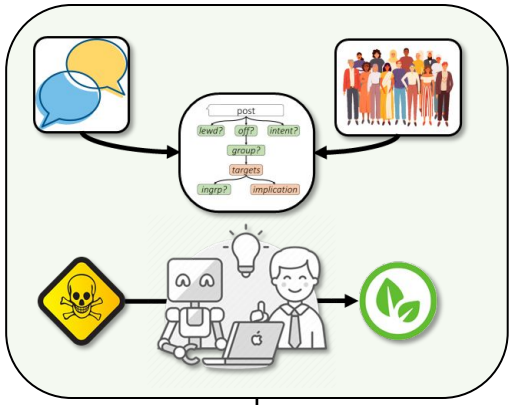


Rewriting and debiasing text



Future work

Human-centric social bias detection and mitigation



Social Bias Frames

New formalism to reason about
social and power implications of language

The content in the rest of this talk may be **upsetting or offensive**

We approach these problems from a **US socio-cultural perspective**

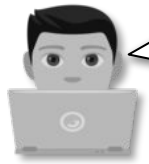
Maarten Sap, Saadia Gabriel, Lianhui Qin,
Dan Jurafsky, Noah A Smith & Yejin Choi

ACL 2020

☆ ***best paper at WeCNLP 2020*** ☆

Two ways harmful social biases are expressed

Overt



We should kill all XYZ

PerspectiveAPI
TOXICITY LEVEL

87%

Social Bias Frames: a new **structured** formalism that distills knowledge about **harmful/biased implications of language**

Subtle

We shouldn't lower our standards just to hire more women.

PerspectiveAPI
TOXICITY LEVEL

8%

women are less qualified than men

We shouldn't lower our standards just to hire more women.

Social Bias Frame

lewdness?

no

offensive?

yes

intentional?

maybe

group or individual?

group

Women

Women are less qualified than men

in-group language?

no

categorical/binary

free-text

Social Bias Frames: motivation

In order to avoid socially biased outputs, models **need to understand what to avoid**
[Pereira et al. '16]

Social Bias Frames: a **new view on social biases**

- **more explainable and trustworthy**, with *explanations* of implied biases [Ribeiro et al., 2016]
- **more holistic** than binary hate speech or offensiveness classification
  - Subjective definitions of hate speech/offensiveness
    [Cowan & Khatchadourian, '03, Ross et al., '17, Waseem '16]
  - Backfires, **biased against minorities**
    [Park and Fung, '17; Dixon et al., '18; ***Sap*** et al. '19]


THE VERGE TECH SCIENCE ENTERTAINMENT MORE

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT


TNW

AI will only succeed when people learn to trust it

by ANNA JOHANSSON — Nov 11, 2018 in CONTRIBUTORS

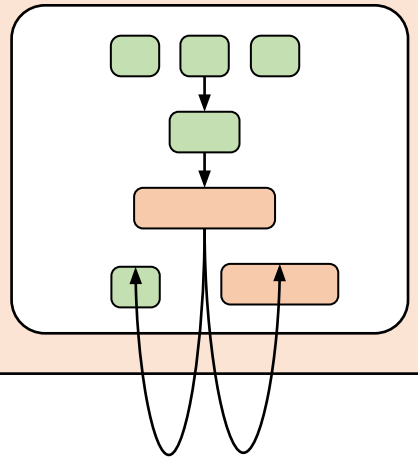

Forbes

6,716 views | Aug 13, 2019, 11:45am

Google's Artificial Intelligence Hate Speech Detector Is 'Racially Biased,' Study Finds

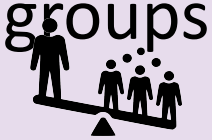
SBIC: Social Bias Inference Corpus

150,000
Social Bias Frame
tuples



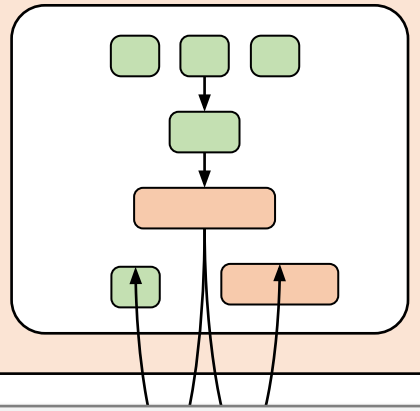
44,000
annotated
(biased) posts

34,000
implications
about 3k
groups



SBIC: Social Bias Inference Corpus

150,000
Social Bias Frame
tuples



44,000
annotated
(biased) posts

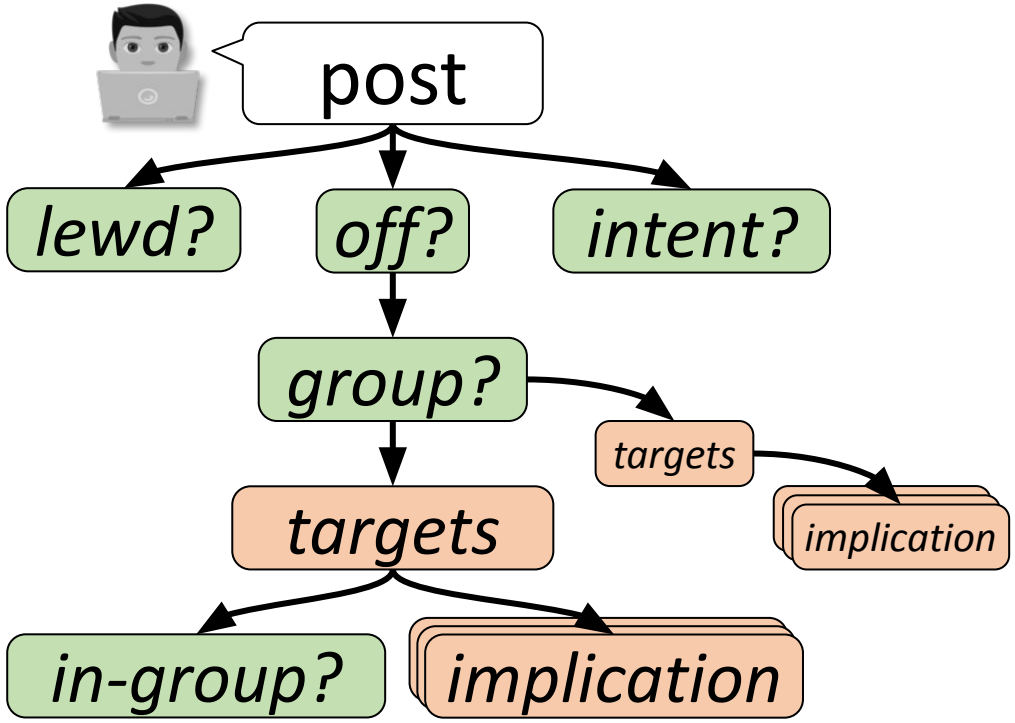
34,000
implications
about 3k
groups

- ***Quality***: trained annotators, 82.4% pairwise agreement
- ***Coverage***: most targeted identities in corpus (women, Black, Muslim) reflects **real-world discrimination** [Robert Wood Johnson Foundation, 2017]

Social Bias Frame design

- Grounded in **social science literature** (rudeness, pragmatics, offense, etc.)
  - Knowing intent can **change perception** of offensiveness, **soften feedback** to author
    [Kasper, '90; Dynel, '15]
  - **In-group statements**, self-deprecation, and reclaimed slurs can appear offensive
    [Greengross & Miller '08]
- **Intersectionality**: posts can target multiple groups, have multiple implied statements
- Pilot studies for high quality annotations

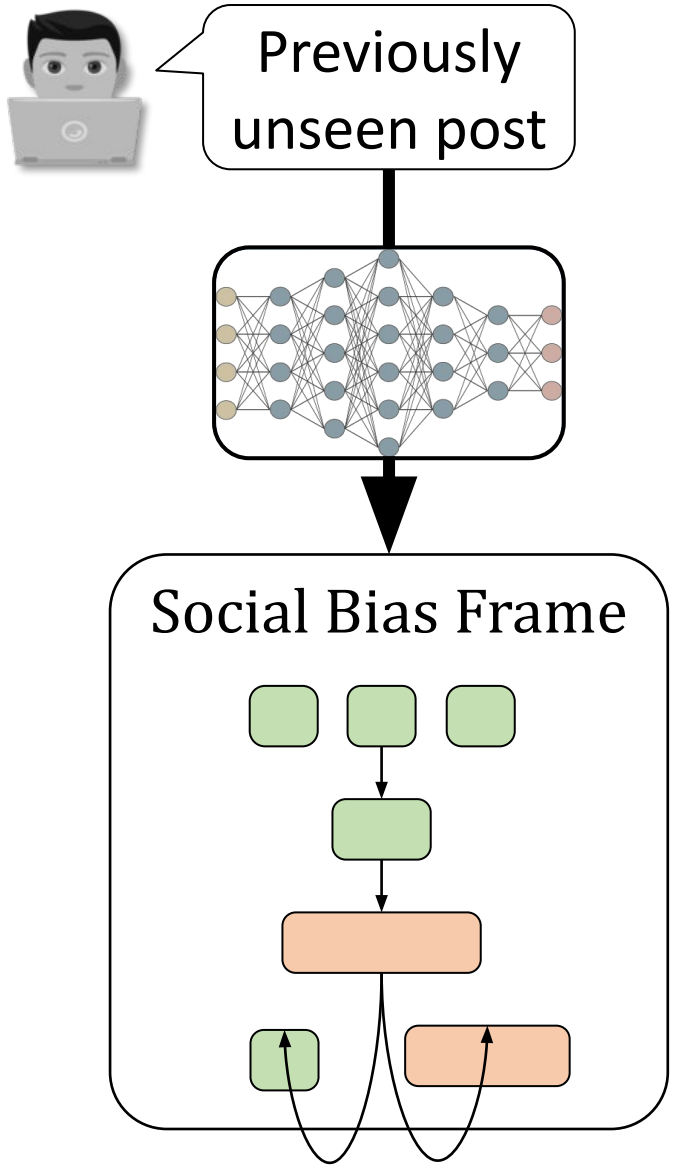
amazon mechanical turk



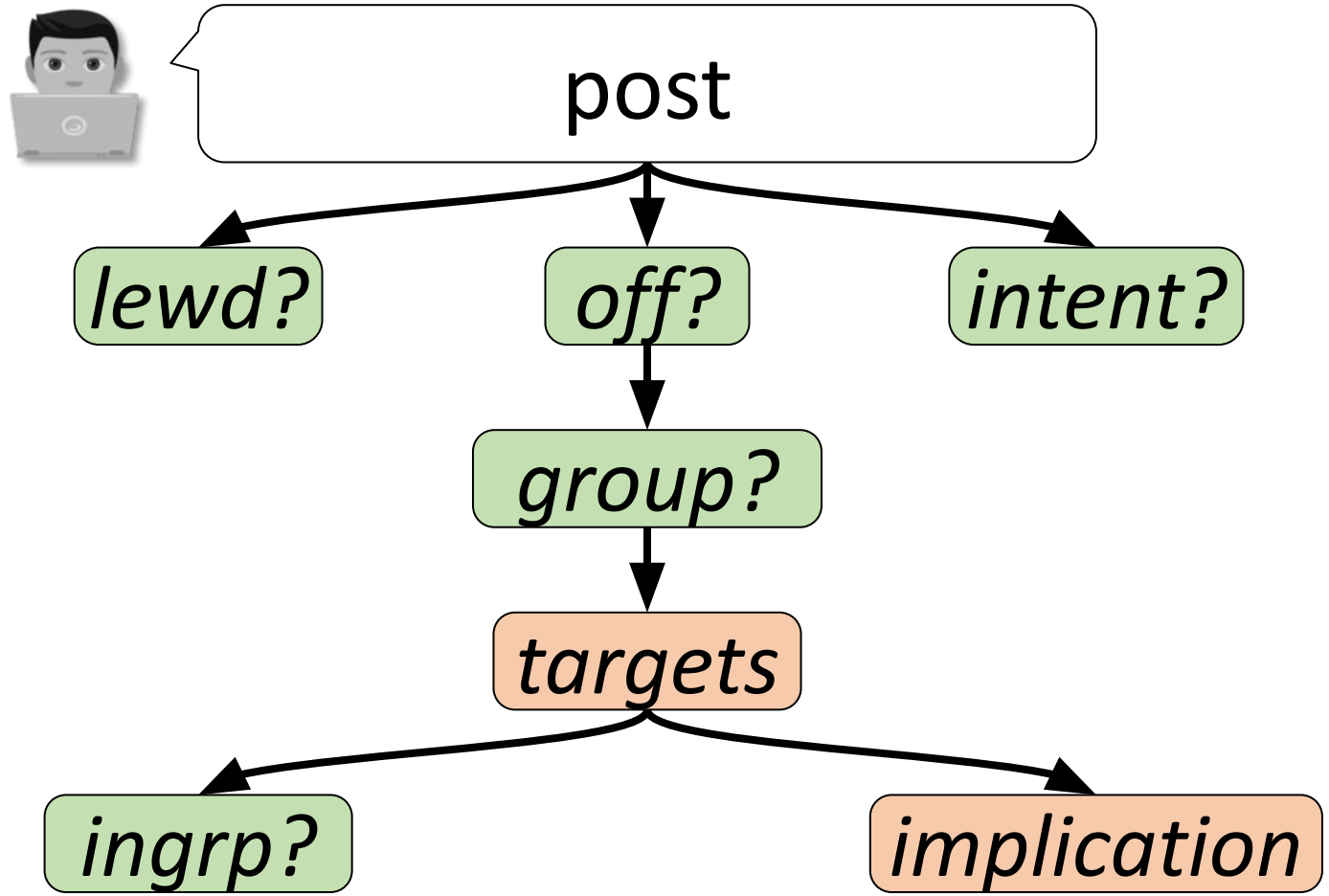
How good are NLP models at making inferences using Social Bias Frames?

Modeling case study

- *Task*: predict entire social bias frame from **previously unseen posts**
- *Requirements*:
  - **Classify** categorical variables
  - **Generate** groups and implied statements
- *Our case study*: GPT-style model

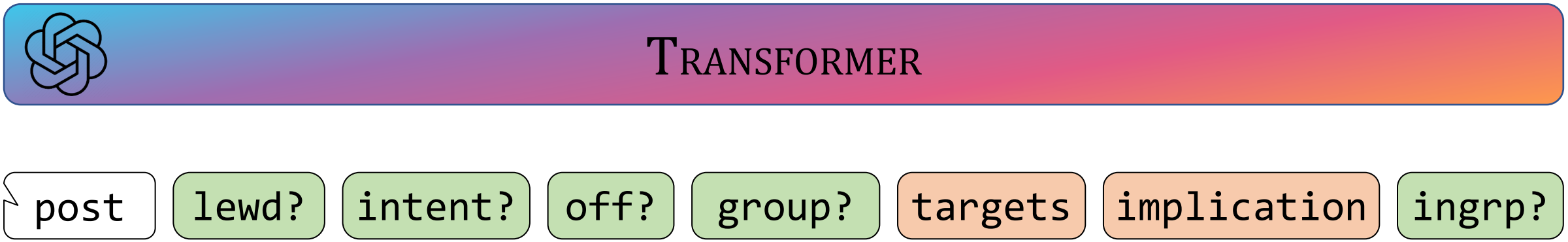


Modelling



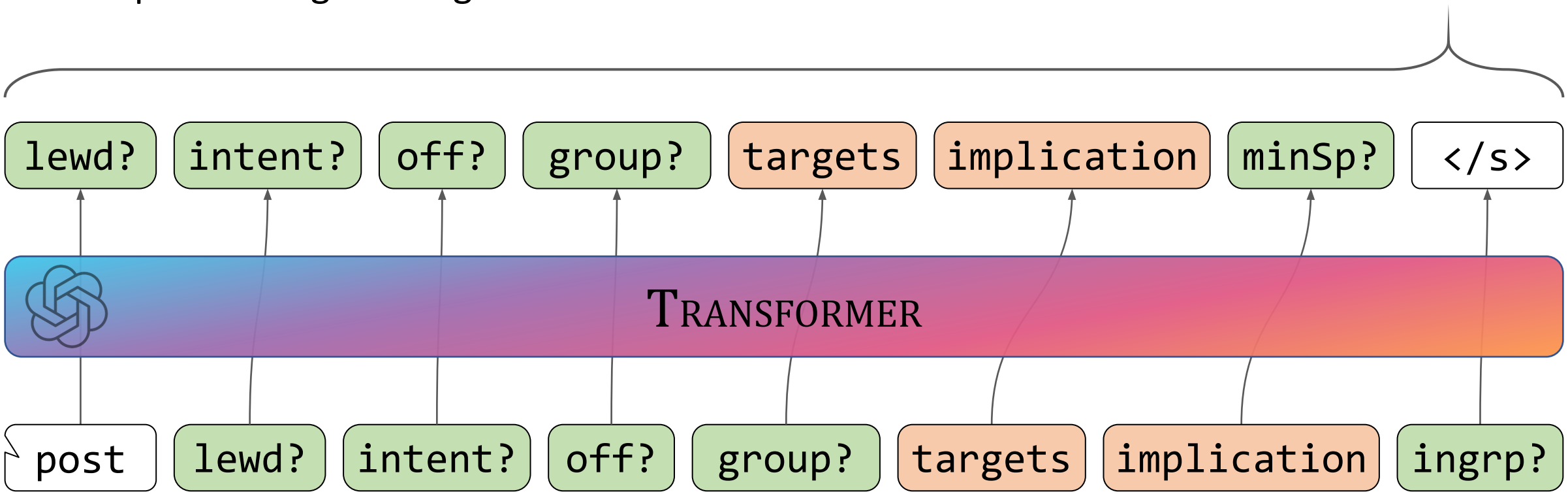
Modelling

- Linearize frame
  - Special tokens for each classification variable



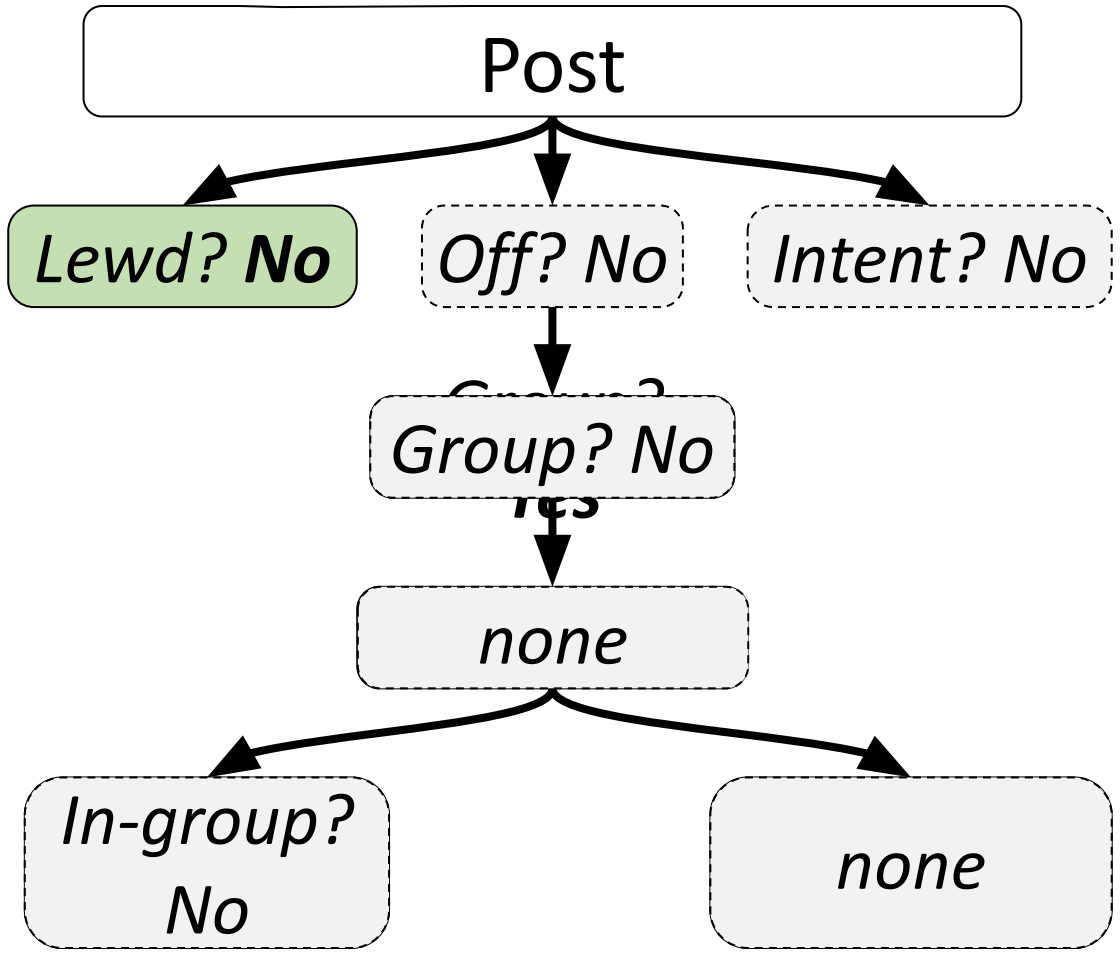
Modelling

- Linearize frame
  - Special tokens for each classification variable
- Transformer-based conditional language model
- Optimize negative log-likelihood of all tokens



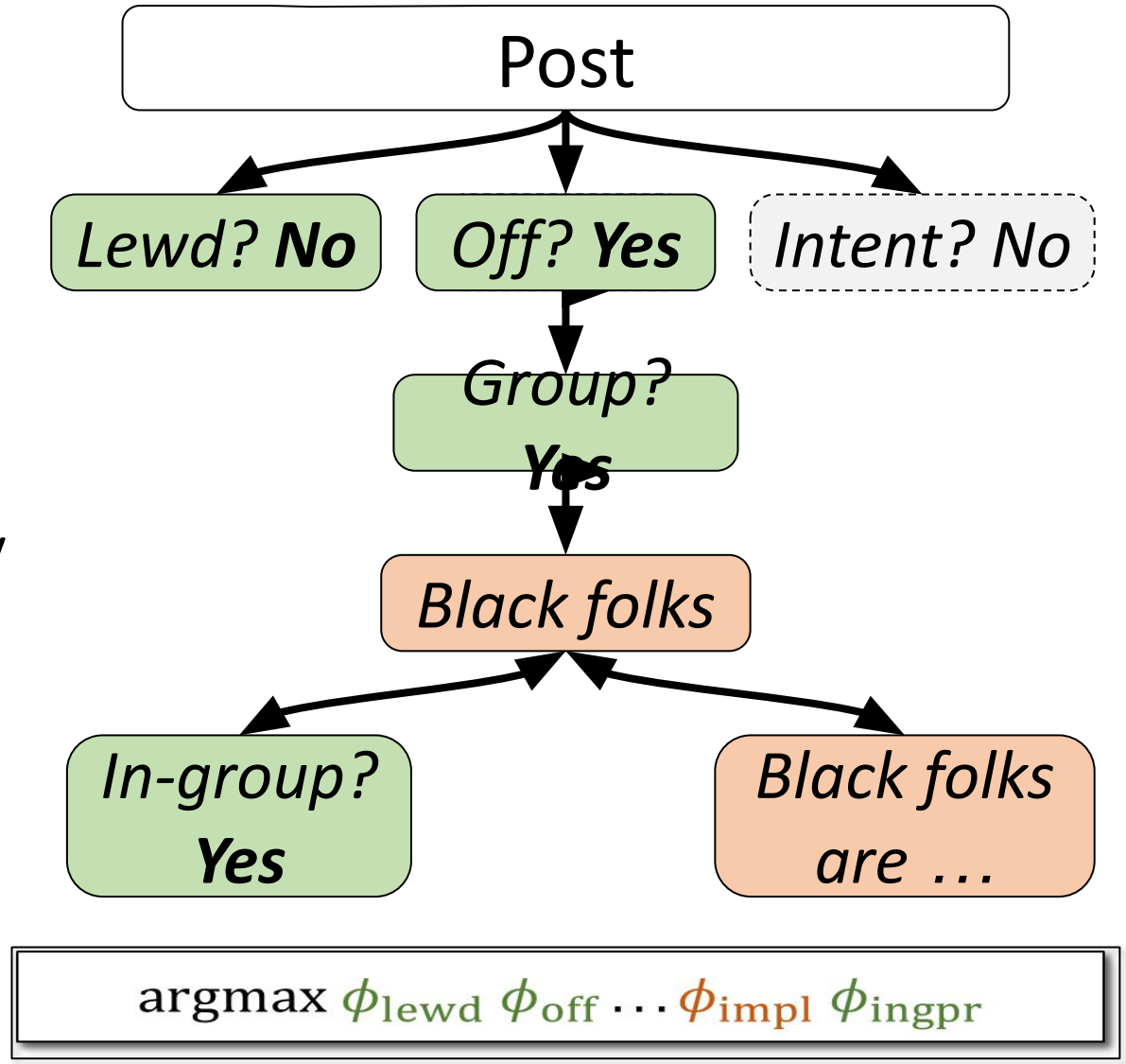
Predicting Social Bias Frames

- Conditional generation of linearized frame, token by token
- **Problem**: generated frame can be inconsistent with frame structure
  - Structure isn't always learned properly by models [Vinyals et al. '14]
- **Fix**: enforce structure post-hoc
  - Naïve inference (top-down)

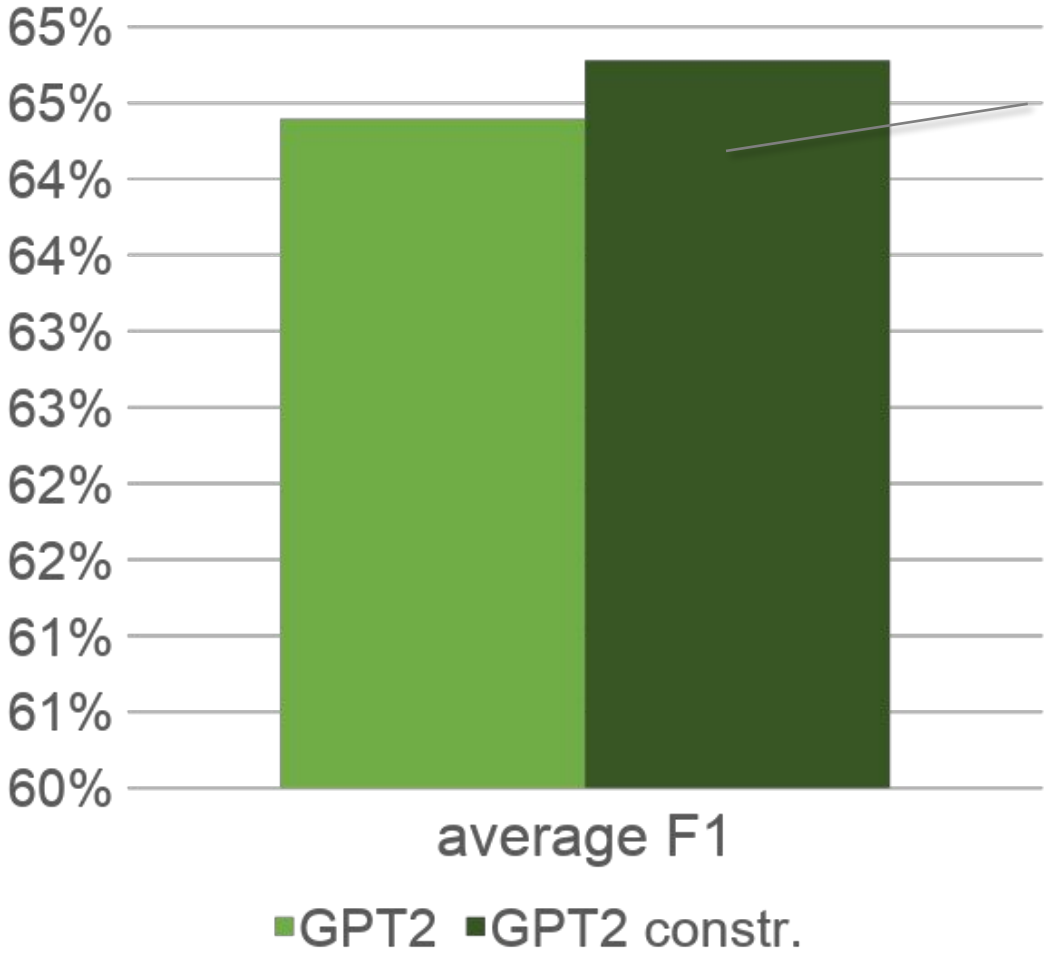


Predicting Social Bias Frames

- Conditional generation of linearized frame, token by token
- **Problem**: generated frame can be inconsistent with frame structure
  - Structure isn't always learned properly by models [Vinyals et al. '14]
- **Fix**: enforce structure post-hoc
  - Naïve inference (top-down)
  - Constrained inference (global)

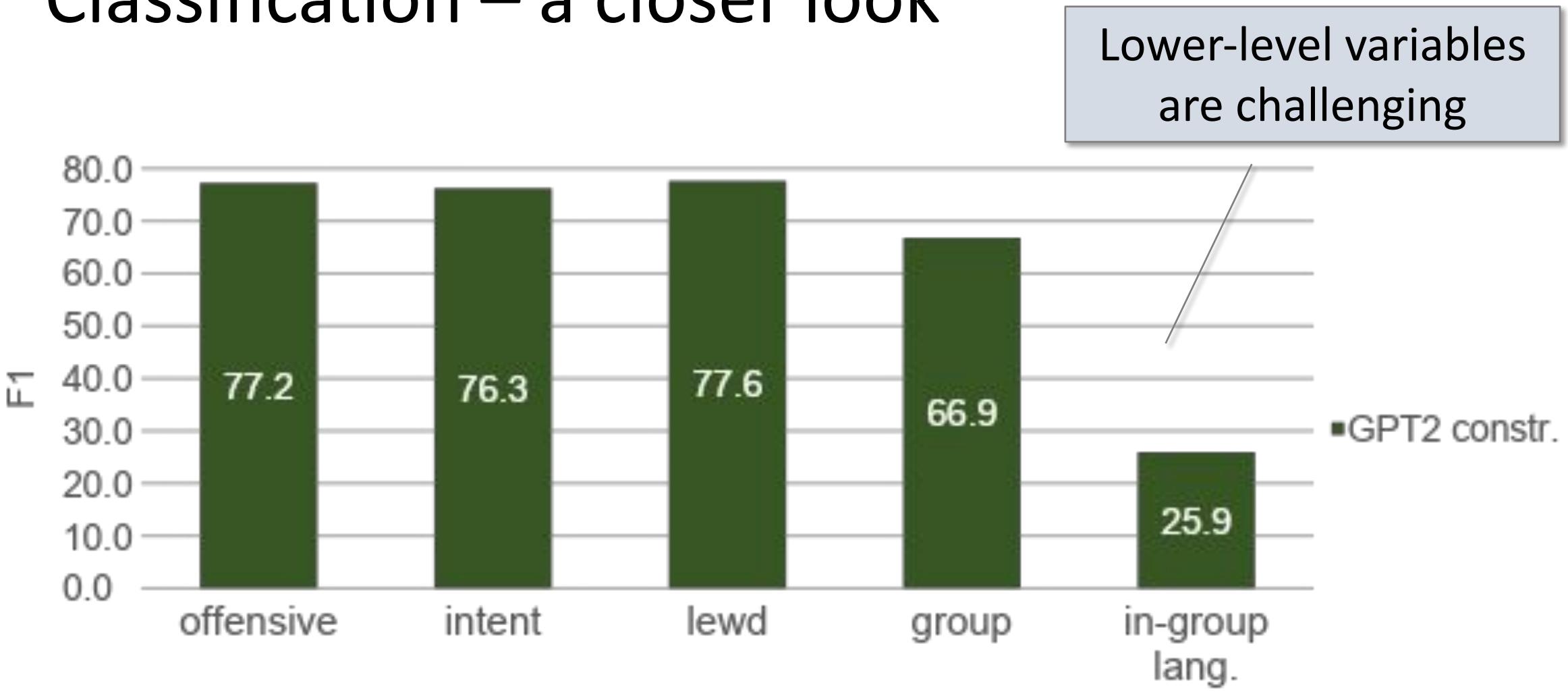


Classification – average performance

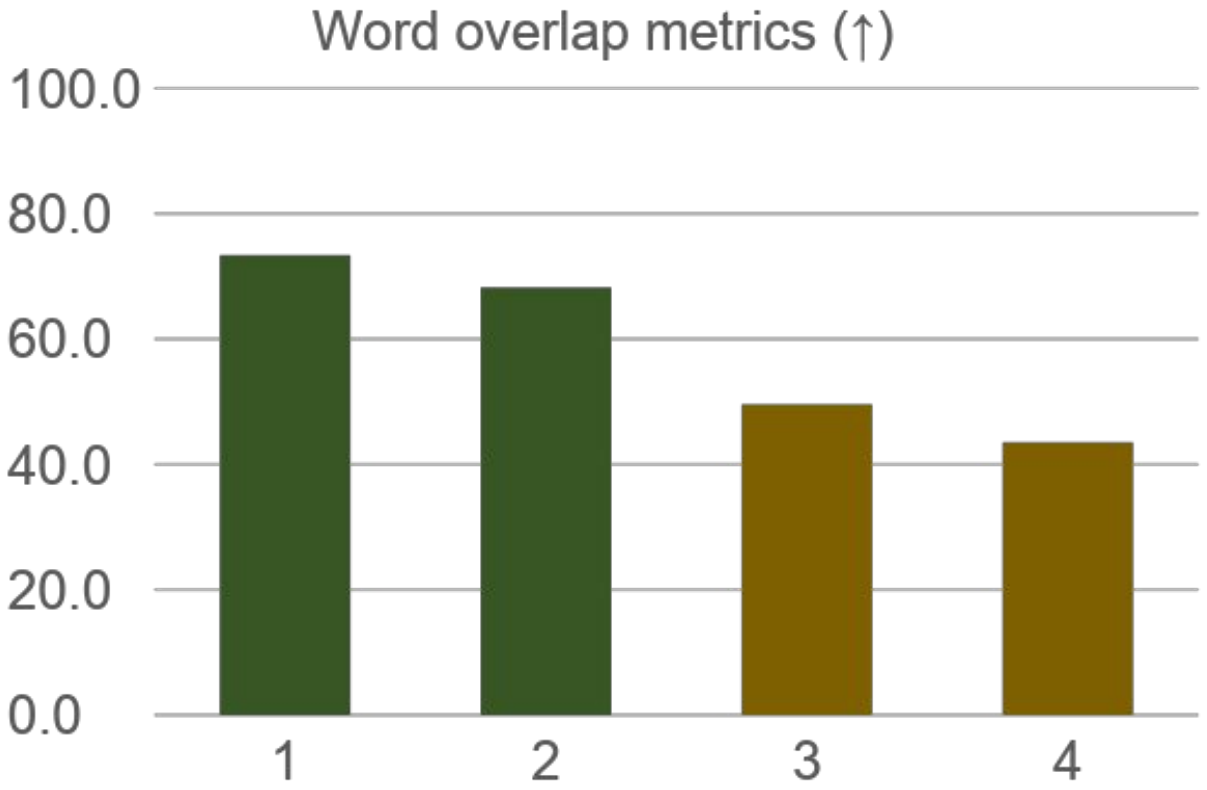


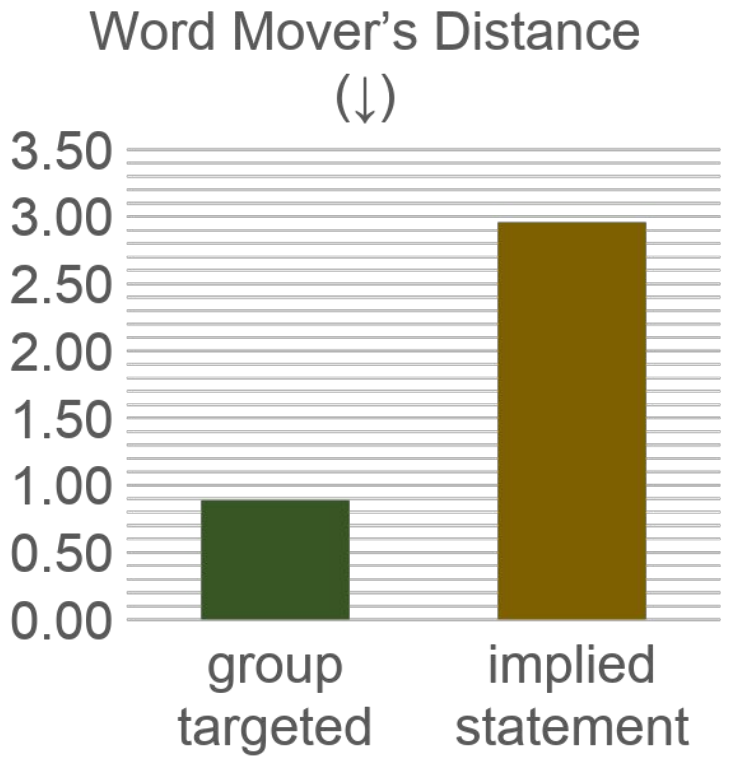
Constrained inference slightly helps

Classification – a closer look



Generation performance





- Model identifies the **right targeted group** relatively well
- The biased **implications more challenging** to generate

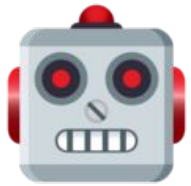
Examples of model predictions

Warning: next slides contain offensive content

Example model inferences

I love gay guys, they're so much fun! I would love to have a gay guy best friend, but lesbians are just annoying.

- Models can be successful for overt biases with verbatim cues
- Struggle with subtle biases, e.g., positive stereotypes [Cheryan & Bodenhausen '00]



*Group? **Yes***

lesbian women

Lesbians are annoying

*Group? **Yes***

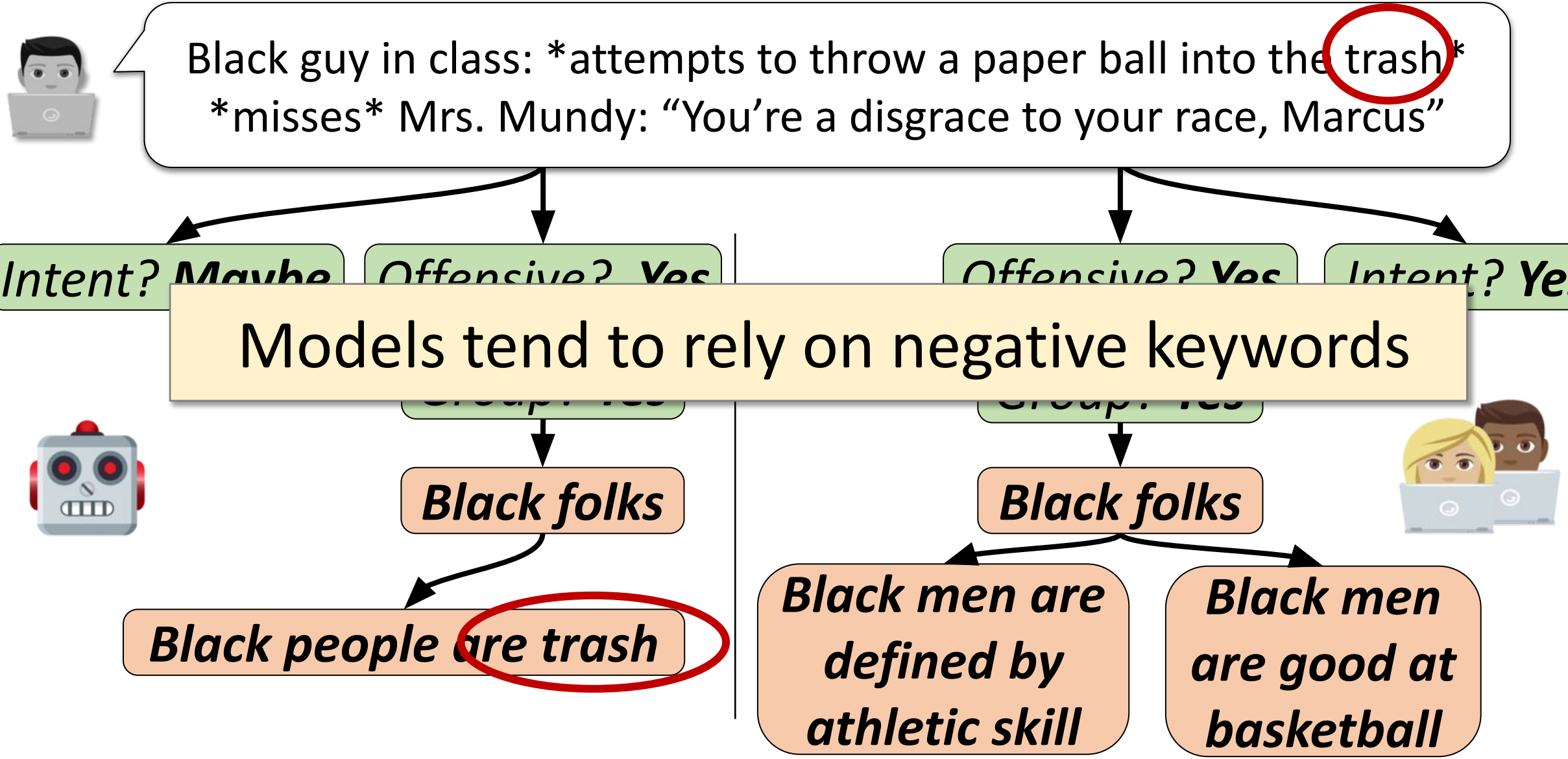
lesbian women

lesbians are annoying

gay men

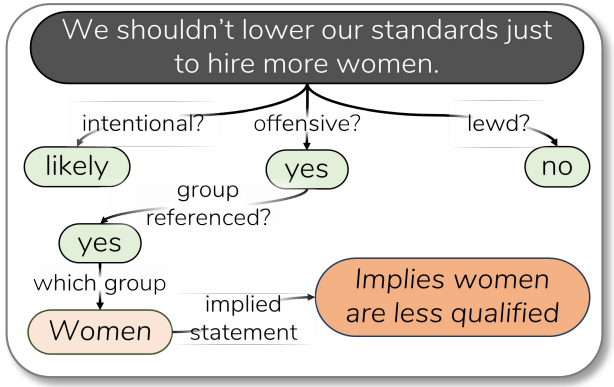
all gay guys are fun to be around

Example model inferences (2)



Social Bias Frames contributions

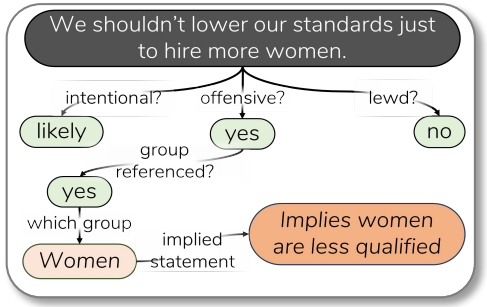
- **New formalism** to distill harmful or biased implications of language
- **New dataset: SBIC**, with 150k annotated tuples
- Experiments showing **models struggle with subtle biased implications**
- Need models that can do better structured reasoning about social biases



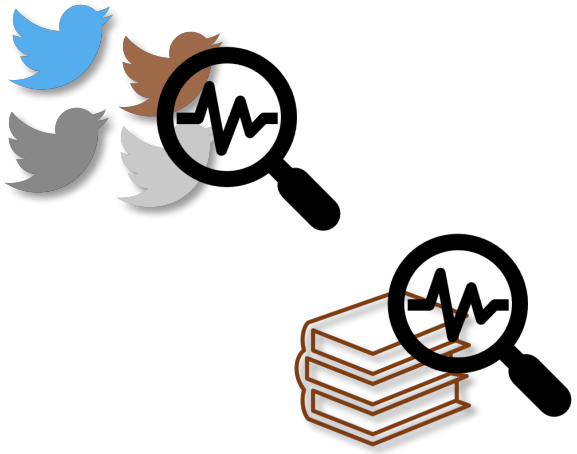
… are just **annoying** …

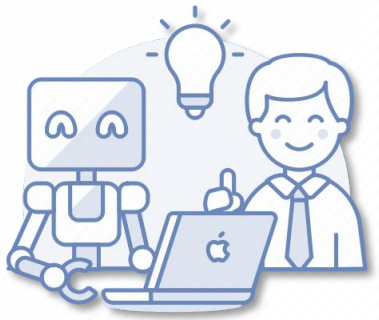
… they're so much fun …

From detecting to rewriting biases



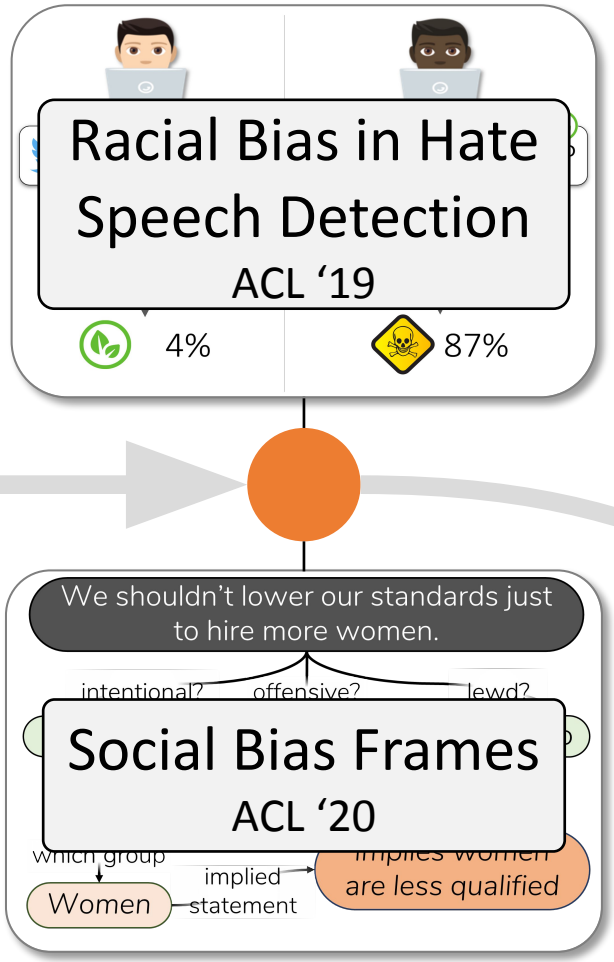
- Social Bias Frames: formalism to **explain social biases** expressed in language
- Explanations can be useful for already written text
  - helping **content moderators** make decisions
  - **quantifying biases** in text corpora
- Explanations can help **authors as they are writing**, by pointing out unintentional biases in their text
- Opens the door for debiasing text through rewriting



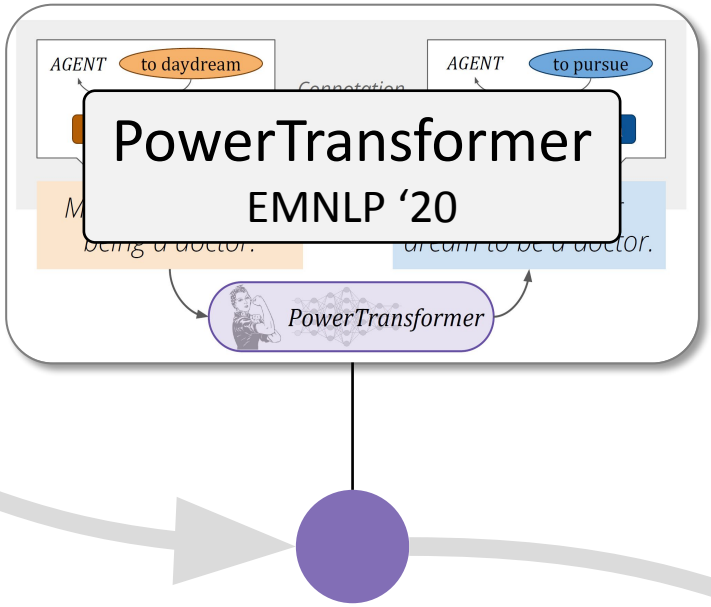


Talk outline

Detecting toxicity and social biases

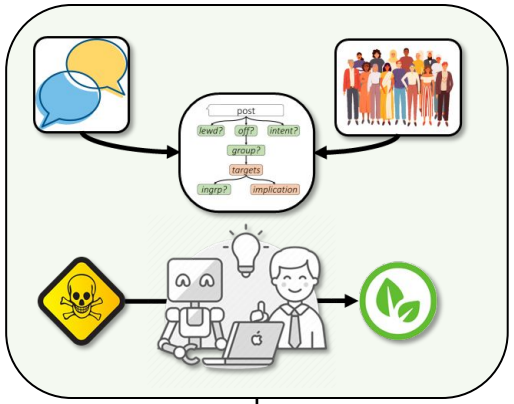


Rewriting and debiasing text



Future work

Human-centric social bias detection and mitigation



PowerTransformer

Unsupervised **controllable revision** model for **biased language correction**

through the lens of *connotation frames* of **power** and **agency**

Xinyao (Michelle) Ma*, **Maarten Sap***,
Hannah Rashkin & Yejin Choi

EMNLP 2020

* Equal contribution



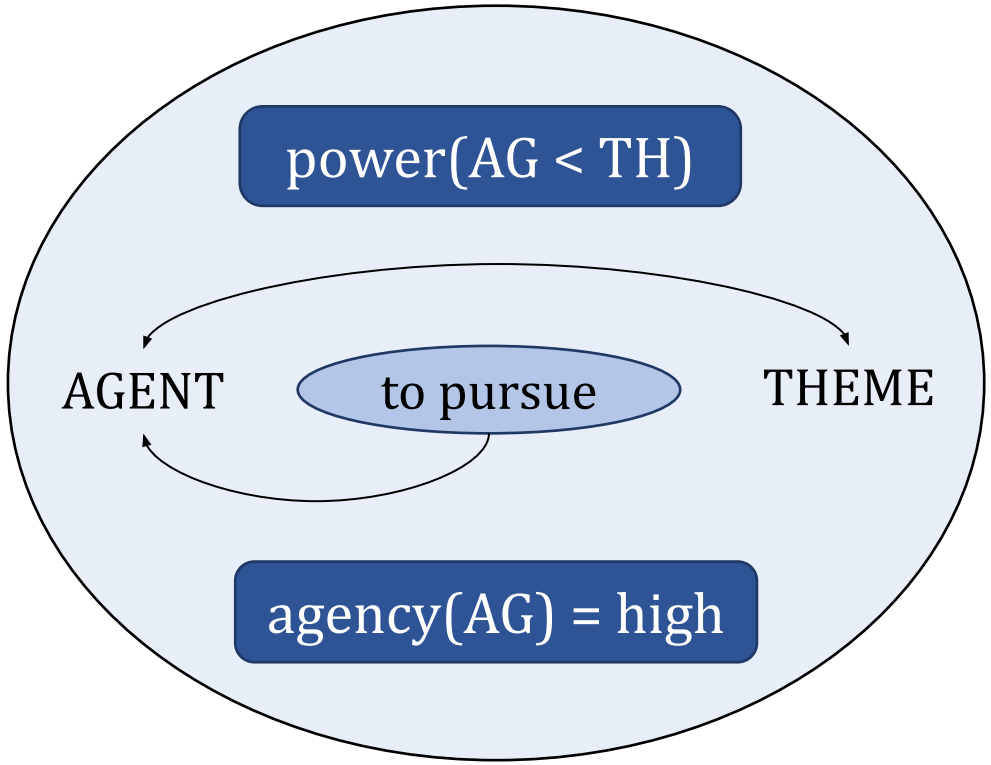
Connotation Frames of Power & Agency [*Sap* et al. '17]

Implied commonsense knowledge around verb predicates

- **Power differentials** between agent and theme of verb
- **Agency** attributed to agent

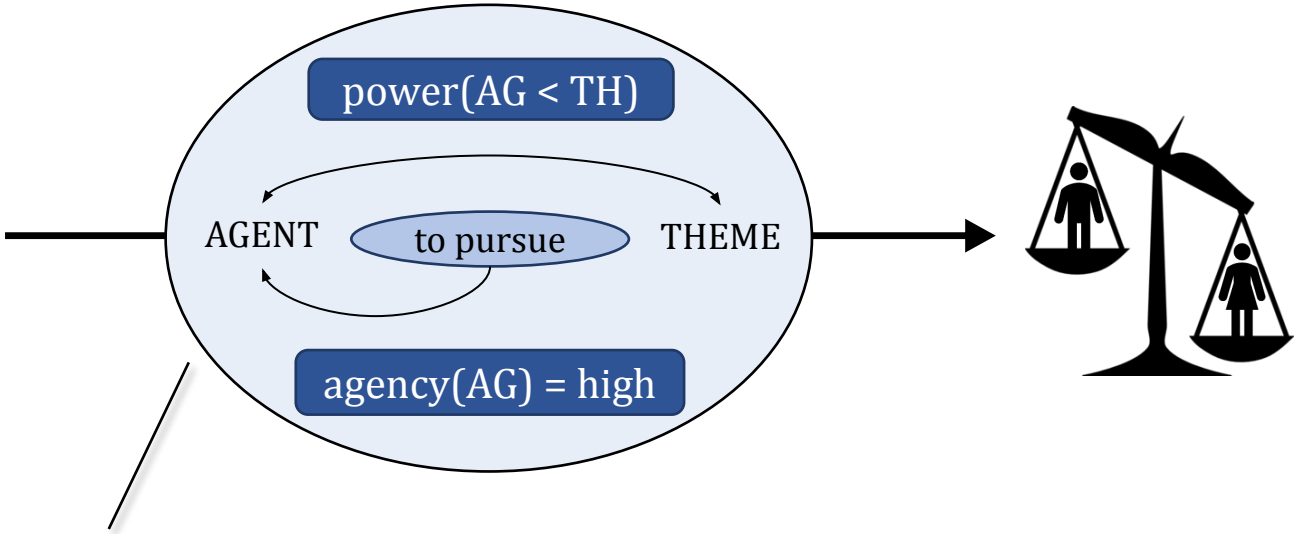
High: decisive, active, drives change

Low: passive, experiences events



Uncovering a new type of gender bias in movies

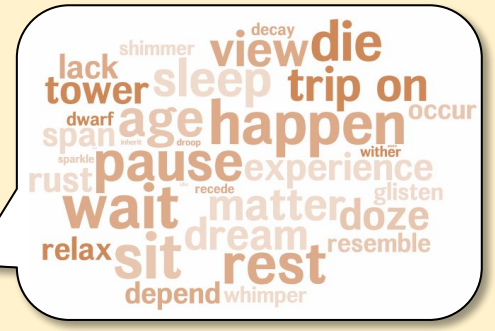
700 modern English movie scripts



Analyses TLDR;

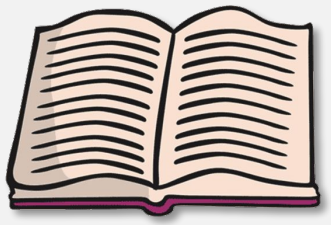
Men: higher agency

Women: lower agency



*See [**Sap** et al. '17] for analyses details*

Cycle of social inequality in text



Text

Women are portrayed as less agentic and less powerful than men

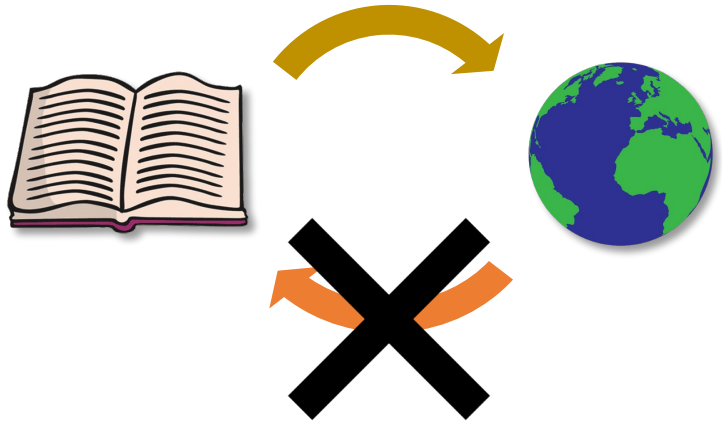
[Lakoff, ‘73; Sap et al, ’17; Field et al. ‘19]

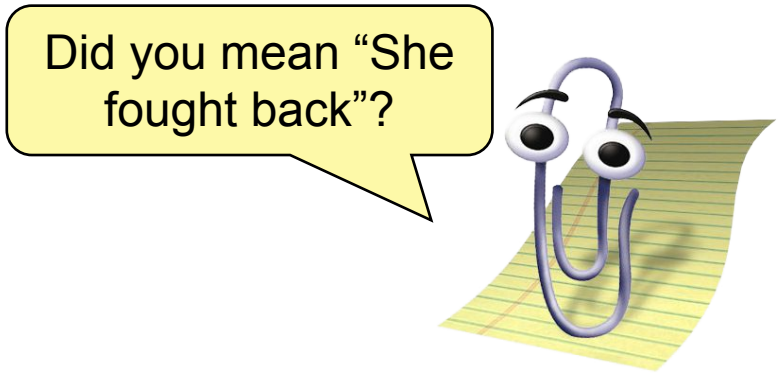
World

Men have more societal and decision-making power than women

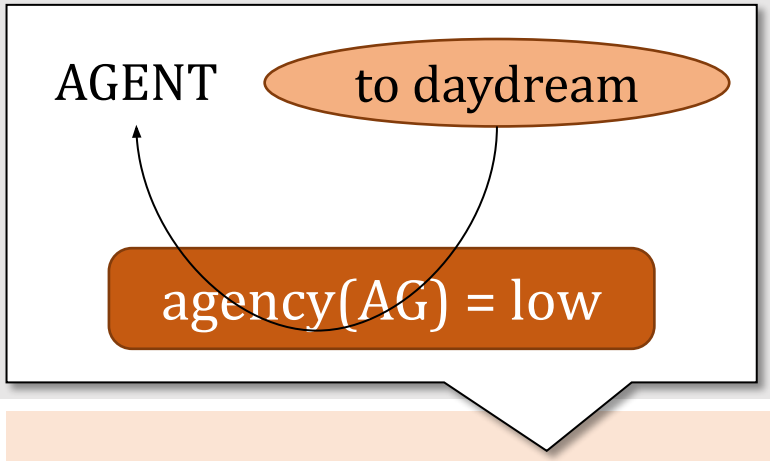
[Peace Corps Report, ‘16]

Controllable debiasing:
can machines learn to revise text to debias portrayals?

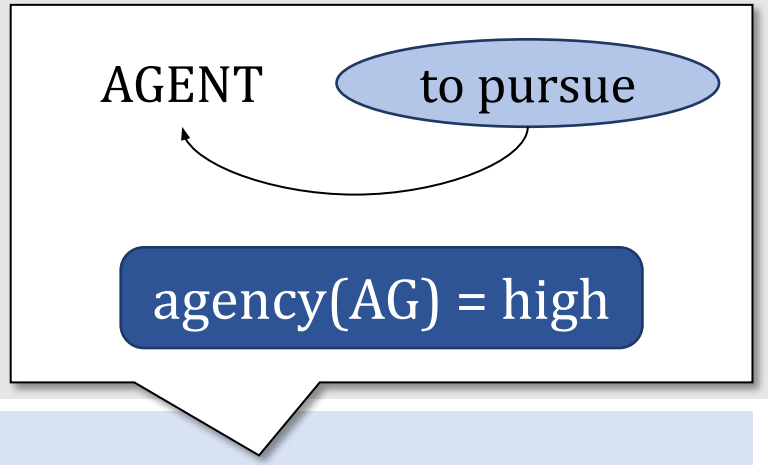




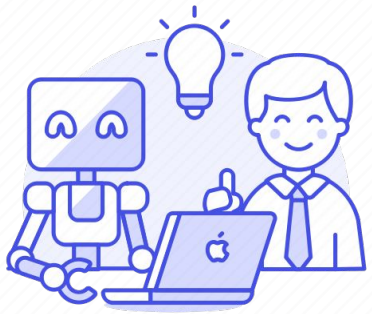
Controllable debiasing of story sentences



Connotation frames of power and agency
*[**Sap** et al., '17]*



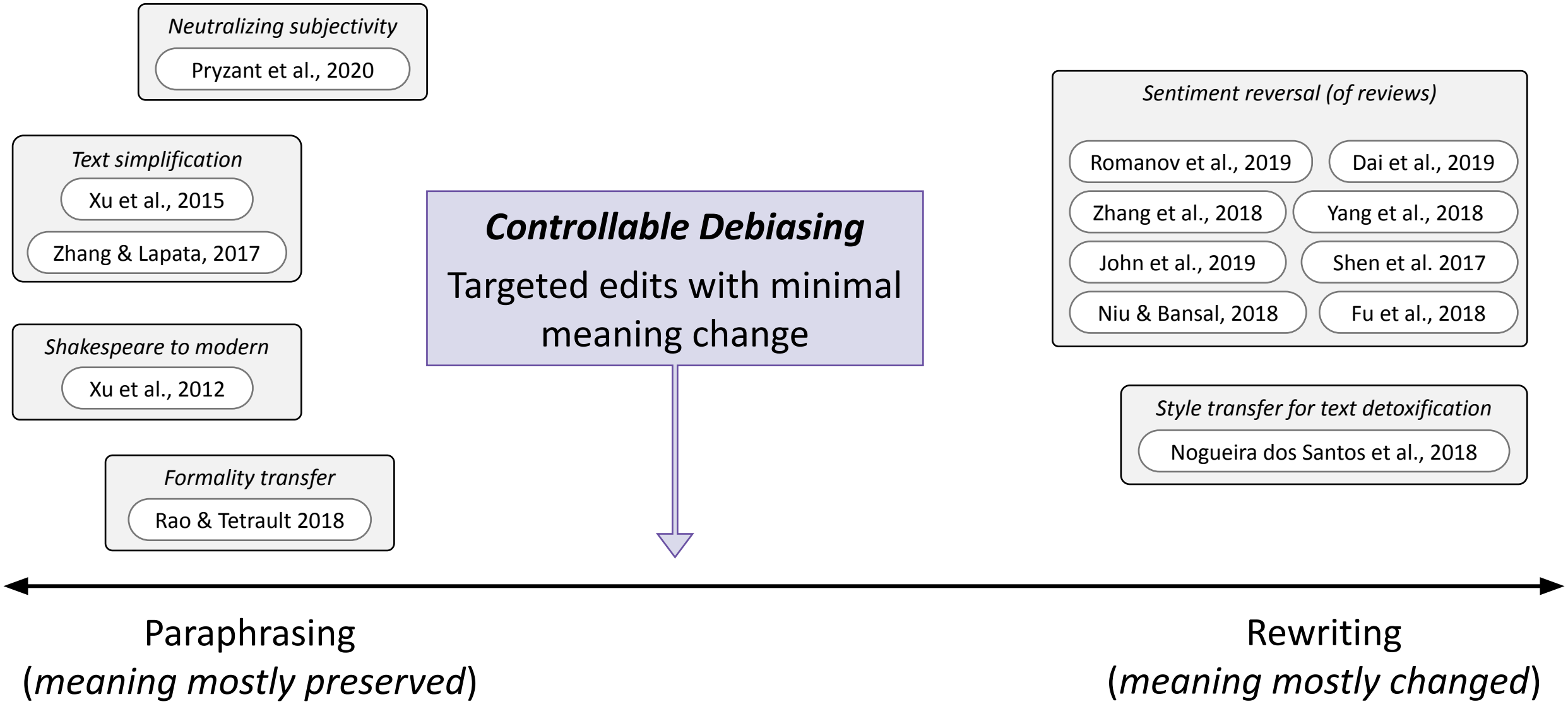
*Mey **daydreams** of being a doctor*



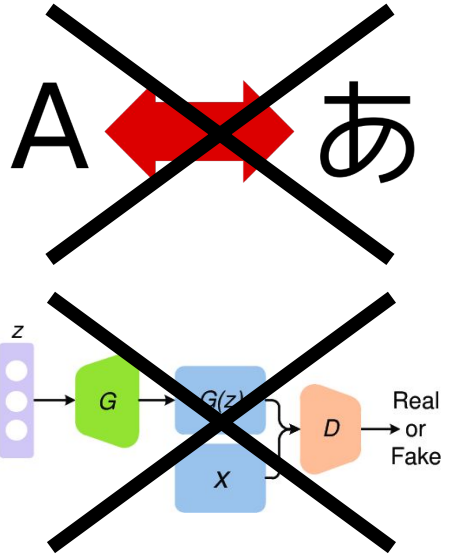
*Mey **pursues** her dream to be a doctor*

PowerTransformer

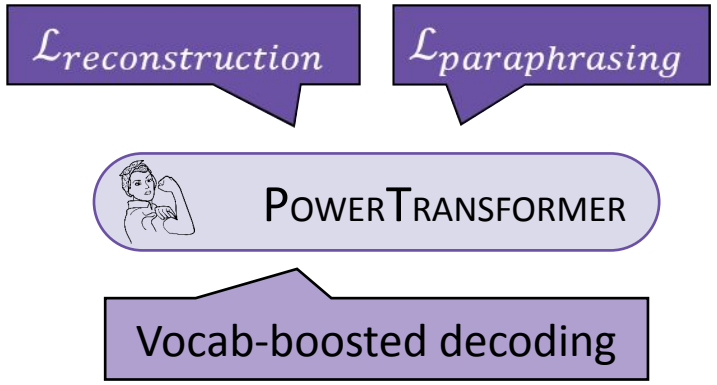
Challenge 1: meaning versus style/framing



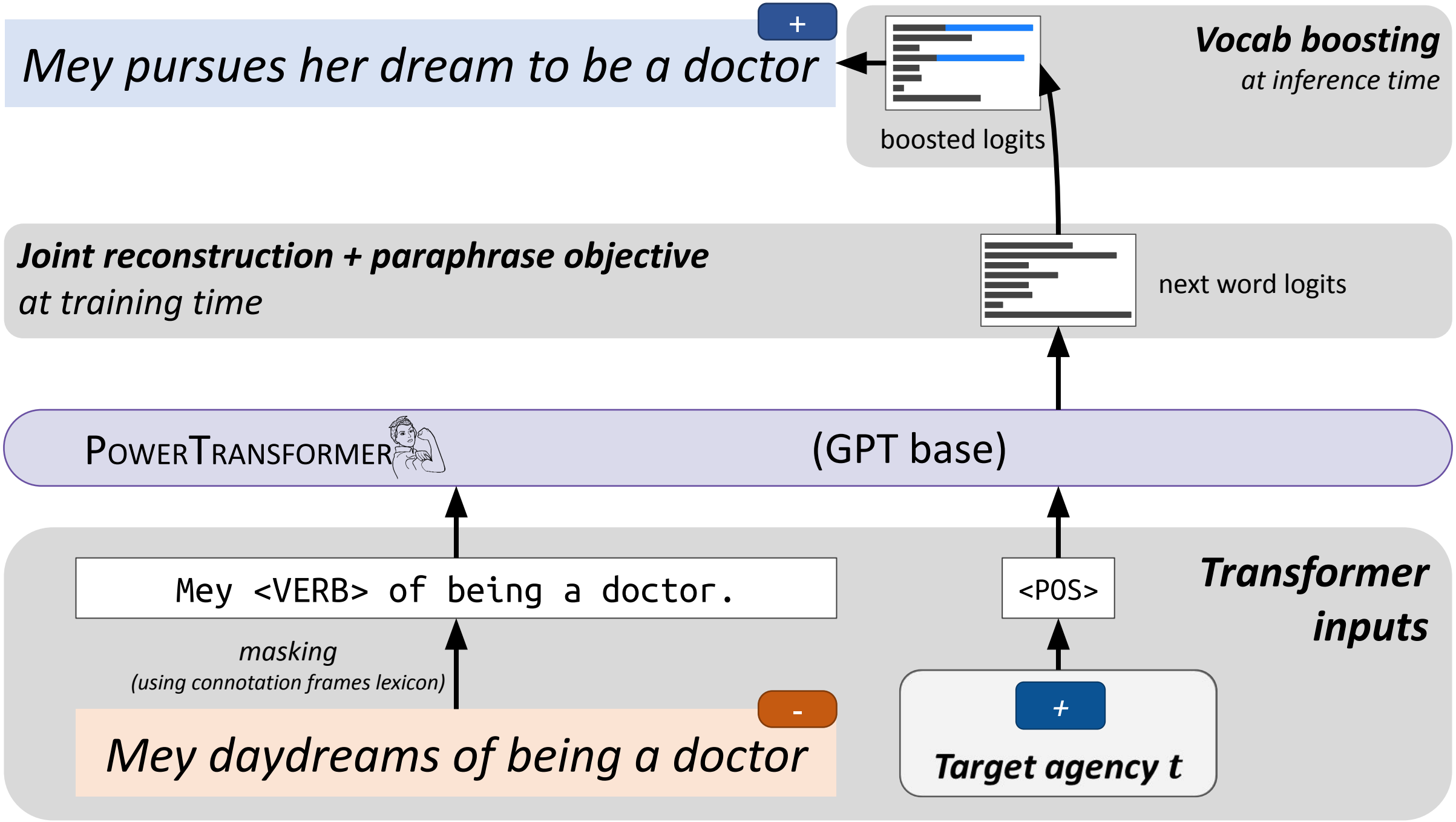
Challenge 2: unsupervised task



- No parallel input-output pairs
  - Machine translation style models won't work
- Generator-discriminators models?
  - No, often leads to less grammatical output text [Li et al. '18]
- ***Our approach***: mask and reconstruct sentences
  - following [Li et al. '18; Sudhakar et al. '19]
- ***Novel modeling aspects***
  - additional paraphrasing training objective
  - vocabulary boosting mechanism at inference time



POWERTRANSFORMER model overview


Mey pursues her dream to be a doctor
+
Vocab boosting
at inference time
boosted logits
Joint reconstruction + paraphrase objective
at training time
next word logits
PowerTransformer
(GPT base)
Mey <VERB> of being a doctor.
masking
(using connotation frames lexicon)
Mey daydreams of being a doctor
-
<POS>
+
Target agency t
Transformer inputs

$$\mathcal{L}_{joint} = \mathcal{L}_{recons} + \mathcal{L}_{para}$$

$\tilde{\mathbf{x}}$: masked input sentence from story corpus [Mostafazadeh et al. '16]

x_i: i-th word in reconstructed sentence
t : target agency level

In-domain objective

$\tilde{\mathbf{y}}$: masked paraphrase sentence from TV subtitle corpus [Creutz 2018]

x_i: i-th word in paraphrase
t : target agency level

Out-of-domain objective

Joint reconstruction + paraphrase objective
at training time

next word logits

POWERTRANSFORMER  (GPT base)

Goal: increase likelihood of tokens that connote target agency

$$p(y_i|y_{<i}, \tilde{\mathbf{x}}, t) = \text{softmax}(Wh_i + \beta \cdot A \cdot t)$$

$A \cdot t$: boosting of target agency tokens

β : boosting strength

Mey pursues her dream to be a doctor

+

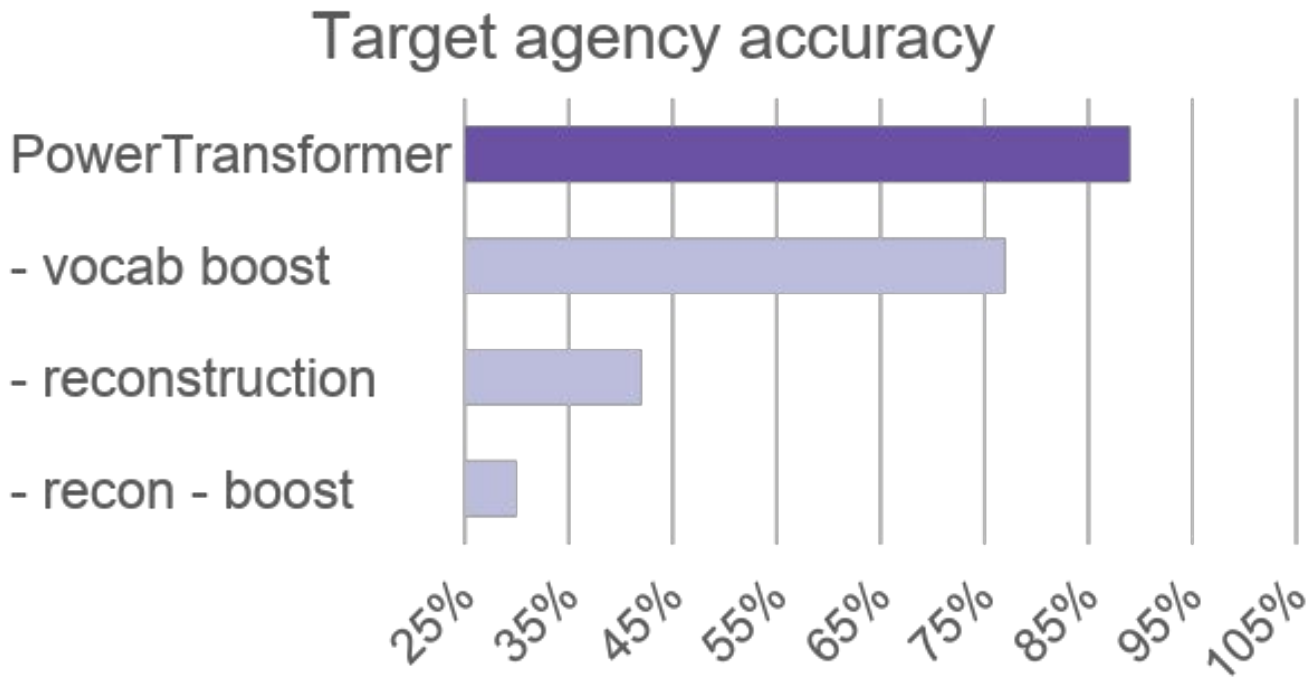
boosted logits

Vocab boosting
at inference time

Joint reconstruction + paraphrase objective
at training time

next word logits

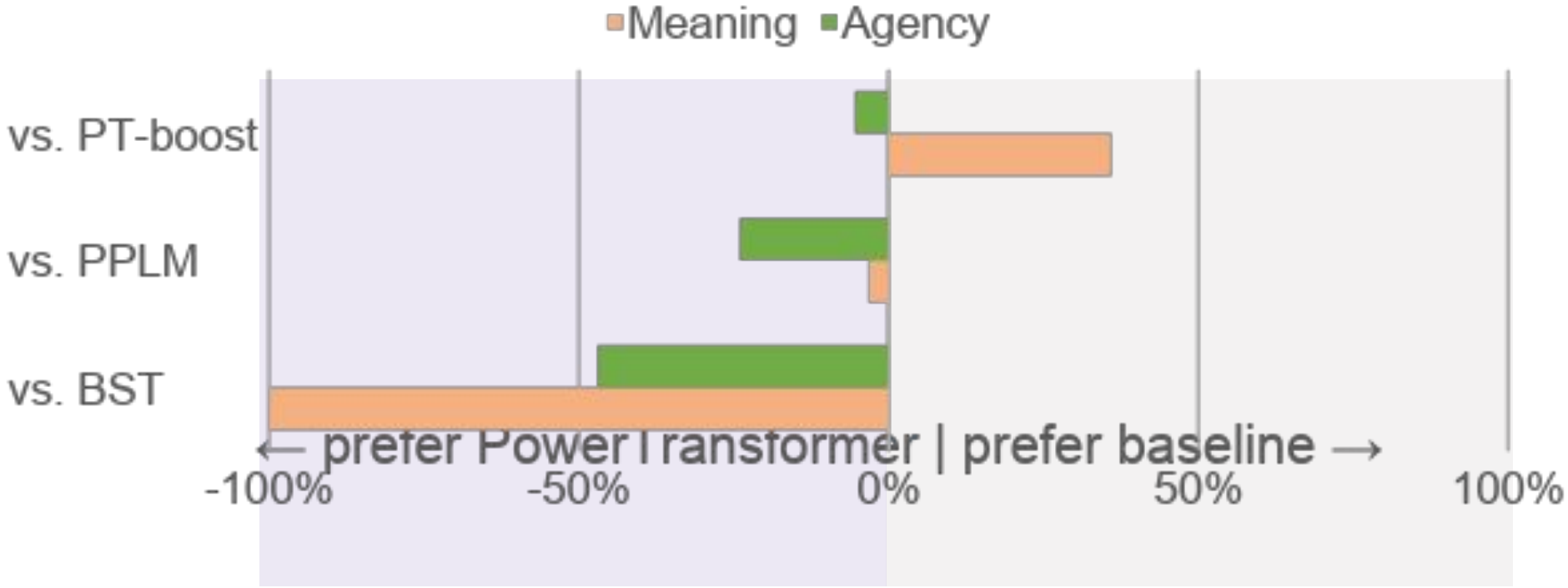
PowerTransformer ablations (automatic eval)



Performance gain from both vocab boosting & joint objectives

Head-to-head **human** evaluation

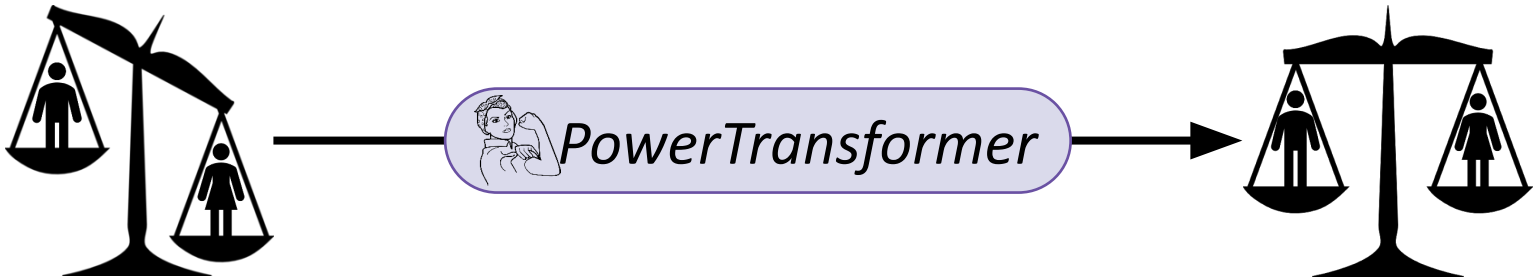
Because automatically evaluating machine generated text is still an open problem [Liu et al. '16; Celikyilmaz et al. '20]



- PowerTransformer models preserve meaning better than previous baselines
- PowerTransformer has more accurate output agency levels

Back to movies:

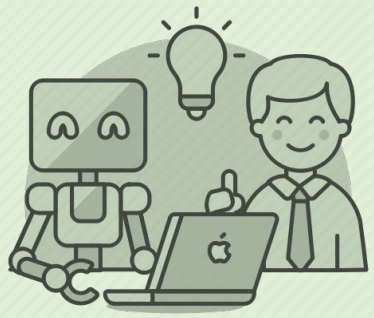
Debiasing character portrayal



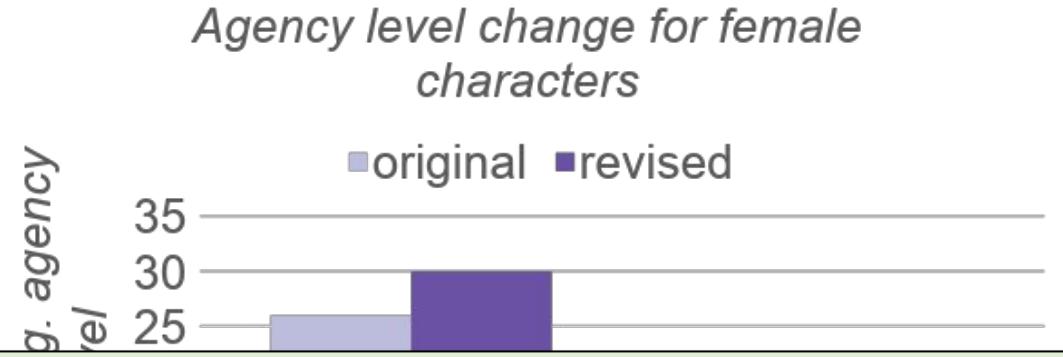
Case study: mitigating gender bias in movie scripts

Original scripts: men portrayed with higher positive agency

Case study: rewrite lines that



Promise for *human-AI collaborative writing* tools to write less stereotypically



agency than before

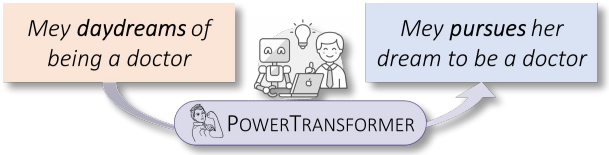
- higher than male characters

original	1.2**	-0.3**
revised	-62.6**	8.7**

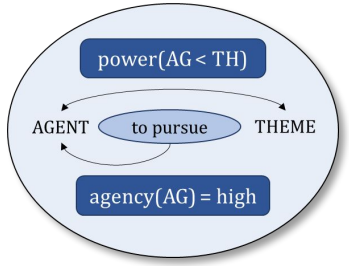
gender effects (β<0: higher for female characters)

**: p<0.001,
logistic regression on gender, controlling for number of words

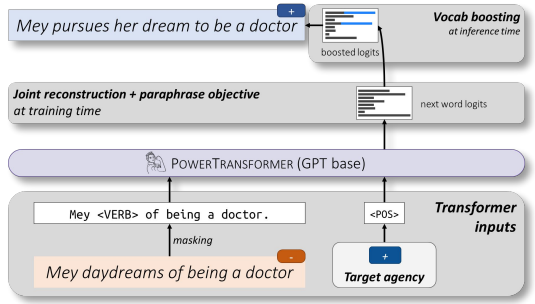
PowerTransformer contributions



- *New task*: **controllable debiasing** of the portrayal of characters in sentences



- *New commonsense formalism:* **connotation frames** of power & agency

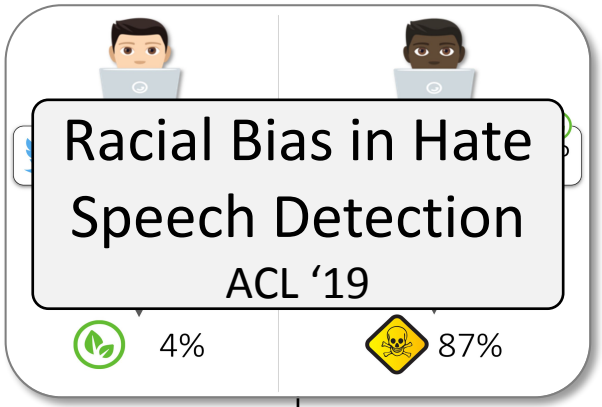


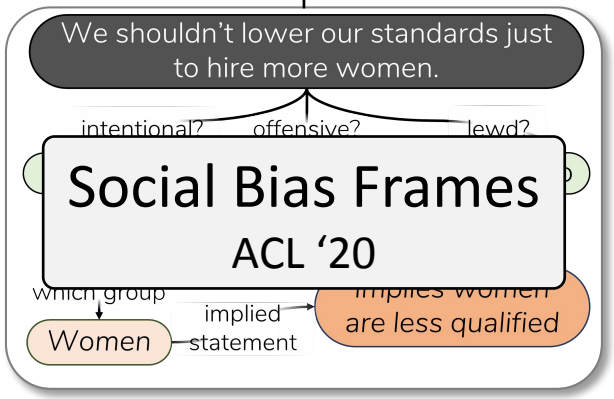
- *New model:* **PowerTransformer:** unsupervised approach to revise sentences using connotation frames

- *Case study*: mitigating gender bias in movies

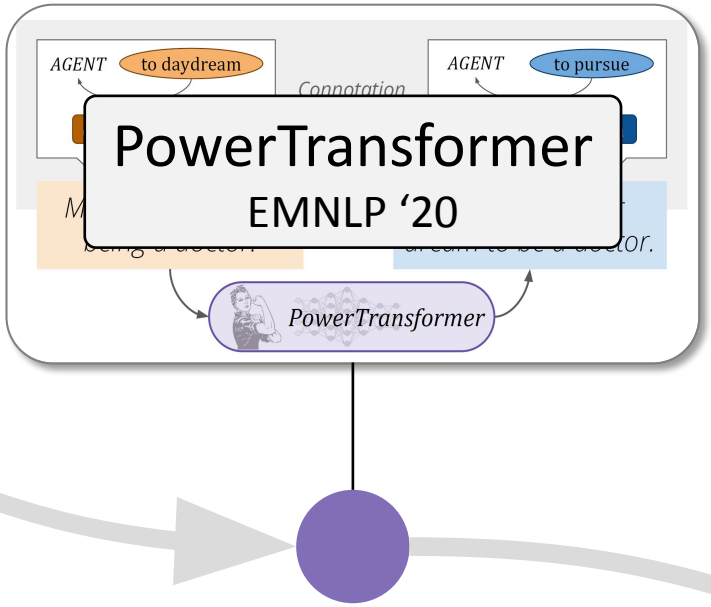
Talk outline

Detecting toxicity and social biases



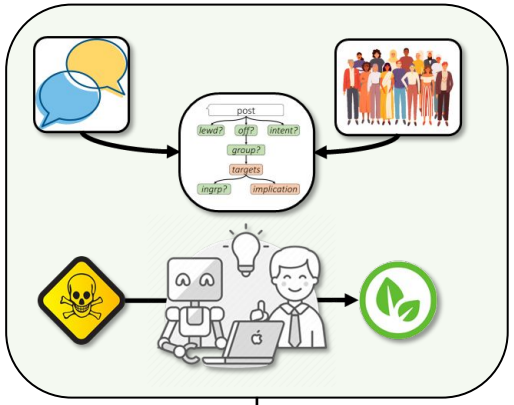


Rewriting and debiasing text



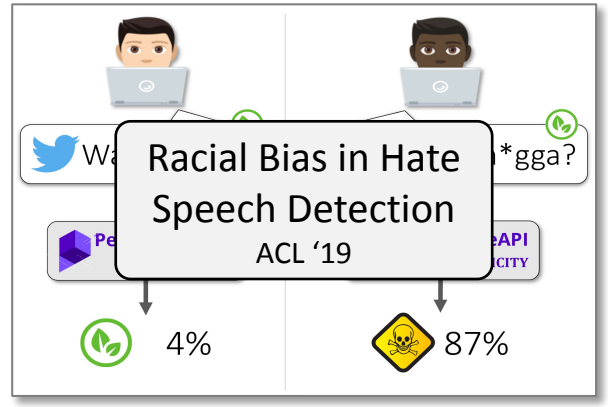
Future work

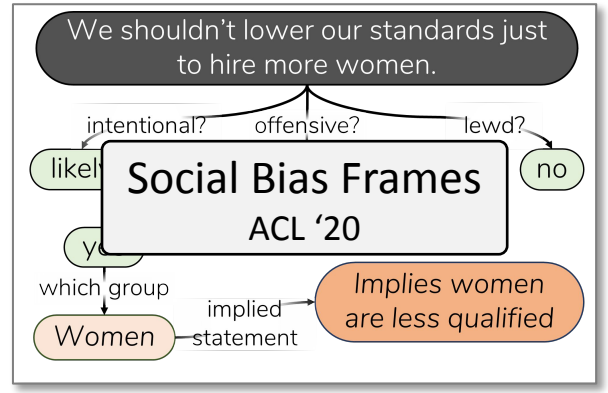
Human-centric social bias detection and mitigation

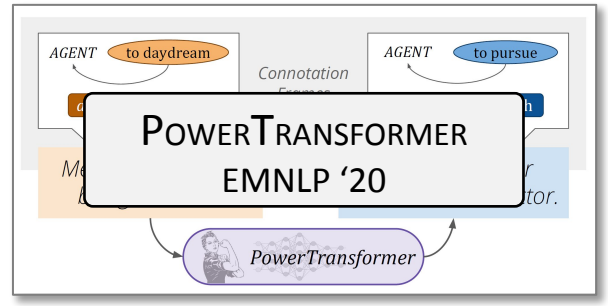


In this talk…

- Uncovered **racial bias in hate speech detection**, proposed race-aware annotation strategies
- Social Bias Frames: new formalism to distill **biased and harmful implications** in language
- PowerTransformer: **revising and debiasing text** through the lens of connotation frames





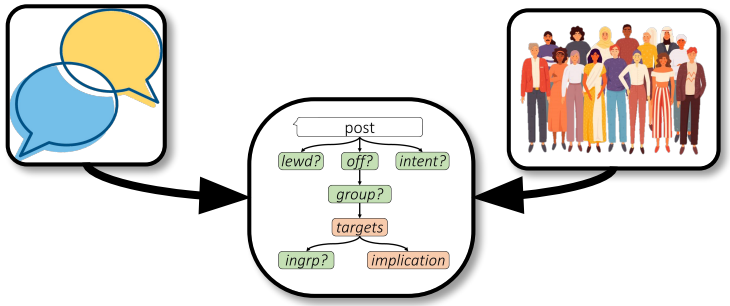


Future directions:
Avoiding and mitigating biases in language with human-centric models

Social biases in human-written language

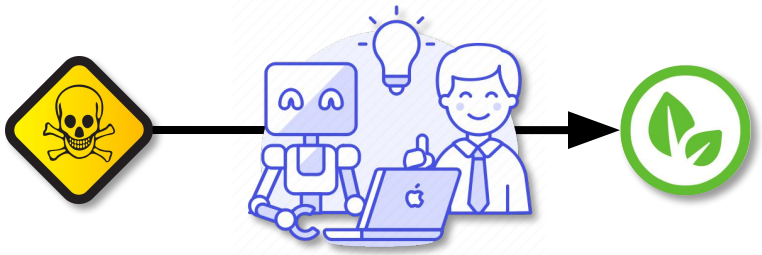
Detecting social biases and toxicity in language

- **New formalisms** for contextual bias representation
  - Incorporate **conversational** context, e.g., using stance towards offensive utterance: ToxiChat [Baheti, Sap, et al., EMNLP '21]
  - Effect of speaker, listener, and annotator **identities**
- **New models** for deeper reasoning about biases in text



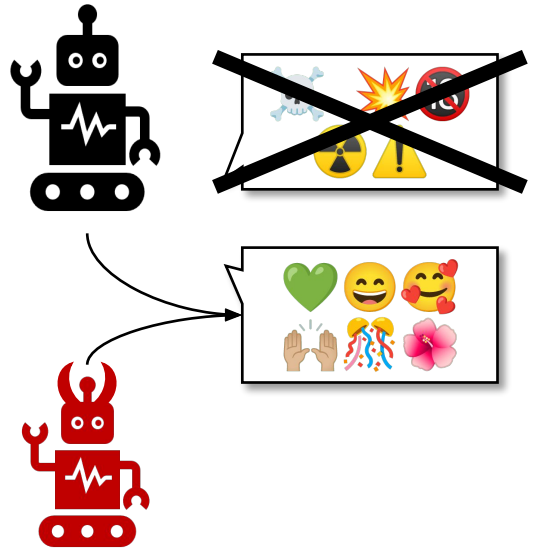
Developing **rewriting systems for debiasing text**

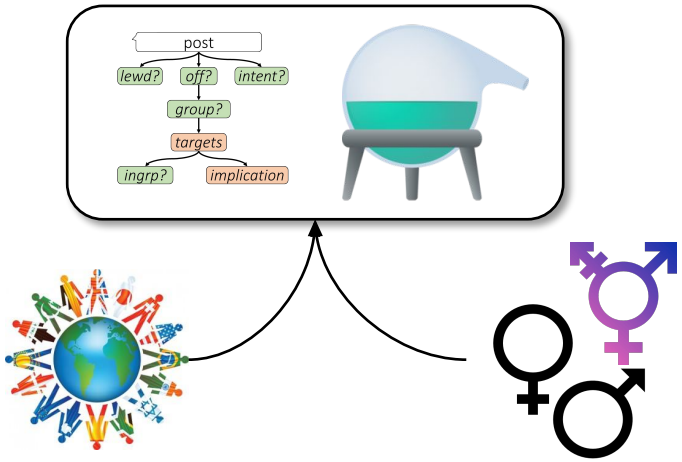
- Using nuanced bias understanding
- With human-AI collaborative writing setups



Avoiding biases and toxicity in machines

- **Keep scrutinizing biases** and toxicity in PTLMs
  - Quantifying neural toxic degeneration
    [Gehman, Gururangan, Sap, et al. '21]
  - Measuring biases in pretraining data
    [Dodge, Sap, et al. '21]
- Develop **steering methods** for avoiding toxicity
  - DExperts: toxic LM as anti-expert for decoding-time controllable text generation [Liu, Sap, et al. '21]
- **Expand dimensions** of what to avoid:
  - **Socially biased**, with Social Bias Frames
  - Violating **social norms**, with Social Chemistry [Forbes et al '21]
  - Allow **personalization** w.r.t. culture and identity





Bridging the gap between NLP and social science

Social sciences for improving NLP

Psychology, sociolinguistics of **perceiving offensiveness in text**

- Effect of annotator demographics, political attitudes, empathy

Mitigating **cognitive biases in crowdsourcing** in NLP tasks

- Effect of task design [***Sap*** et al. '19]

NLP for social science questions

Machine inference about **misinformation in news**

- Automatically distill common tropes evoked by headlines
- Studying fear-inducing framings

Creating methods for analyzing social phenomena in text

- Power & agency in movies [***Sap*** et al. '17]

UNIVERSITY of WASHINGTON

(select) undergrad/Master students

Thanks to my collaborators