

NLP, Ethics, and Society

Warning: this talk contains content that
could be upsetting or offensive



Language Technologies Institute

Carnegie Mellon University

1950s

1980s

2010s

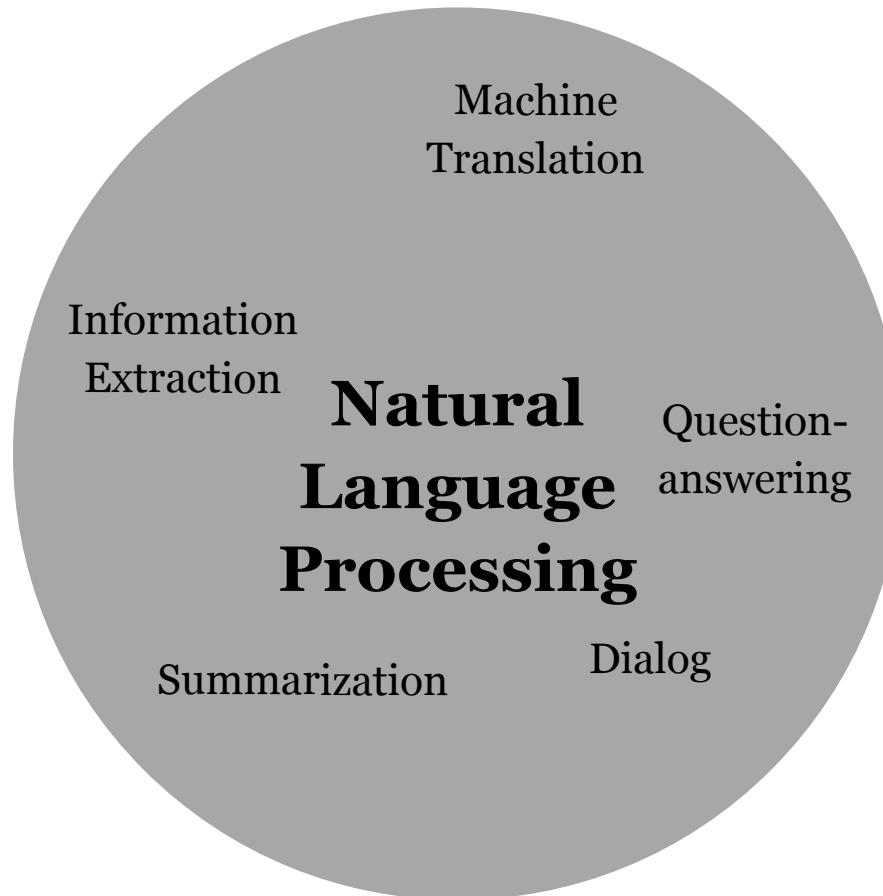
2020s

Rule-based

Statistical Models

Neural Networks

Pre-trained
Language
Models



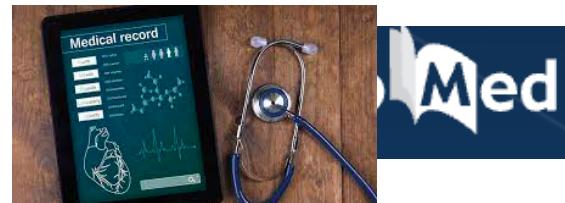
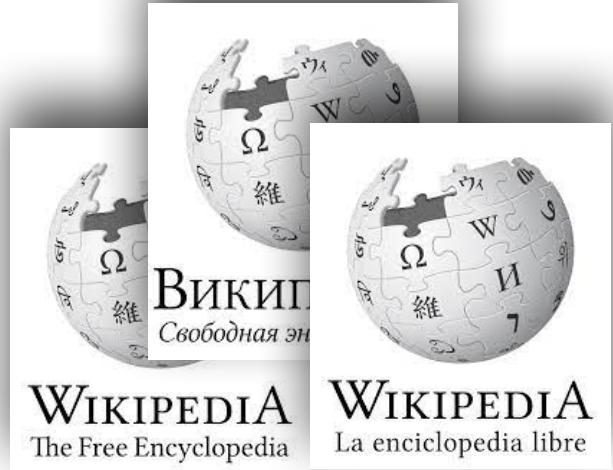
Social Media

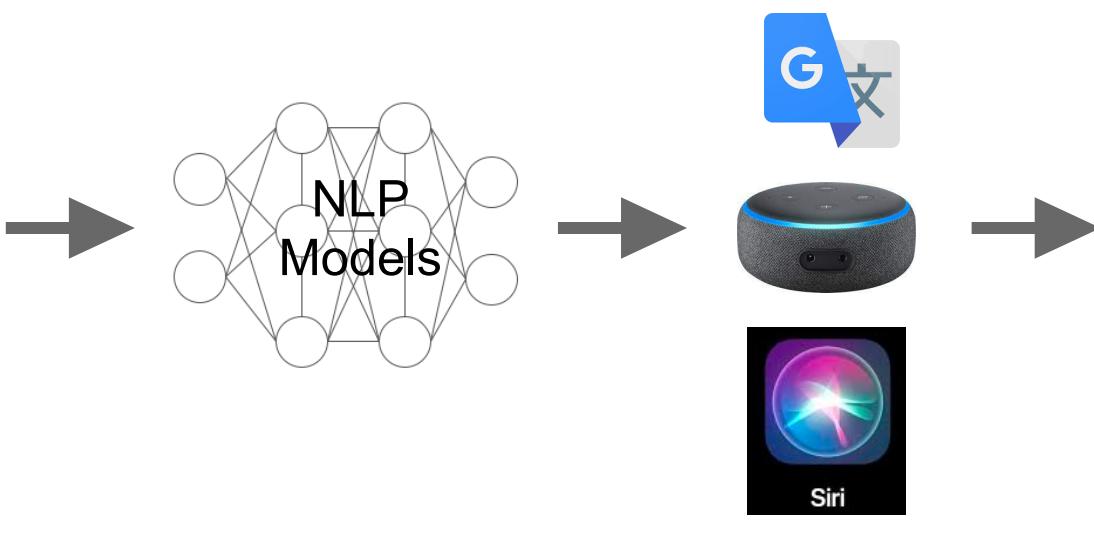


News and blogs

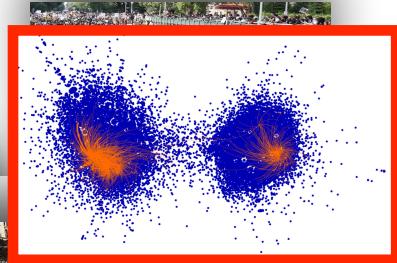
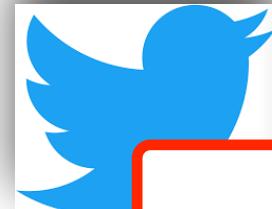


Encyclopedias, text books, and expert notes





Toxicity



You f**** a***



Bias, Stereotypes, and Prejudice Manipulation

WEBMD HEALTH NEWS
**Most Americans
Have Been Duped by
Misinformation**

AllSides™ Media Bias Chart™

All ratings are based on online content only — not TV, print, or radio content.
Ratings do not reflect accuracy or credibility; they reflect perspective only.



ALTERNET
BuzzFeedNEWS

abc
AP politics &
fact check

Ap

The American

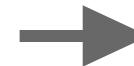
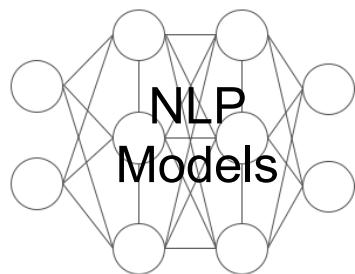
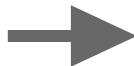
THE AMERICAN
SPECTATOR

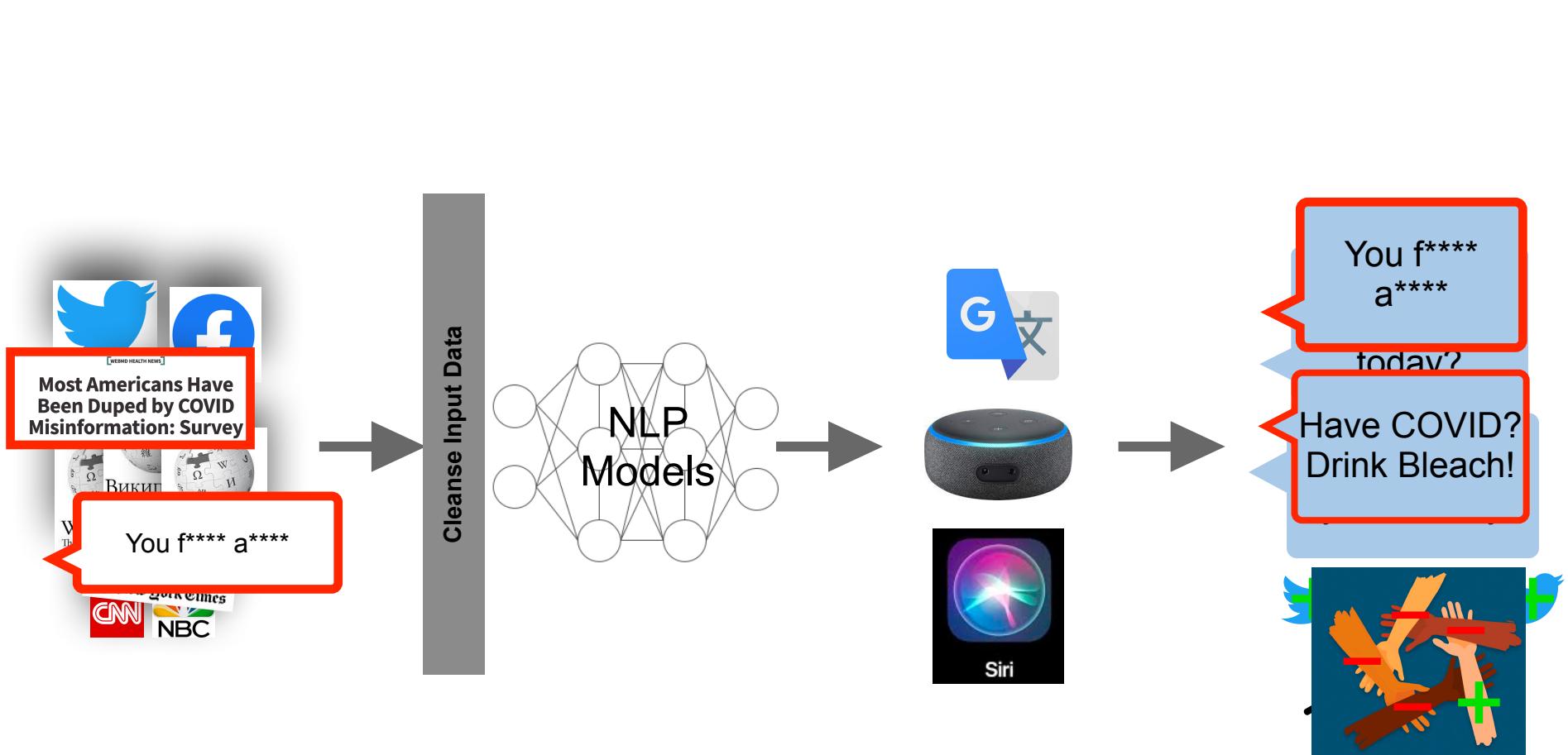
It's a Man's Wikipedia?
Assessing Gender Inequality in an Online Encyclopedia

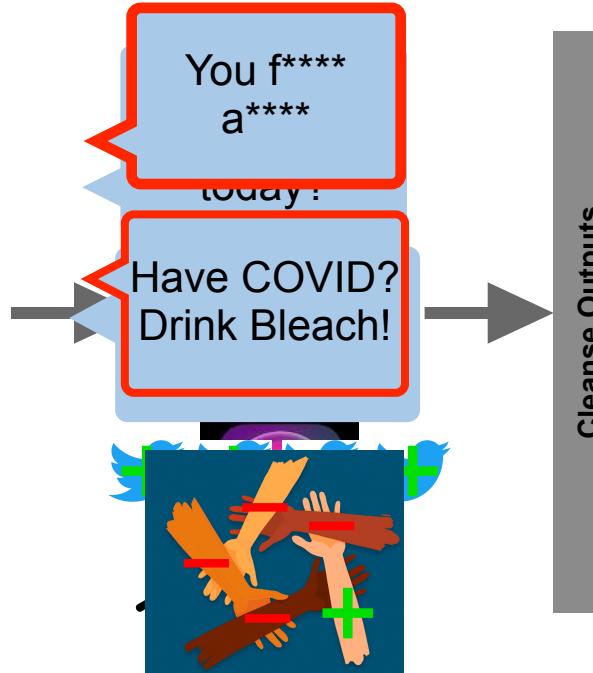
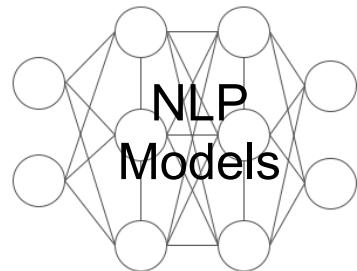
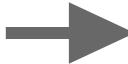
Bias,

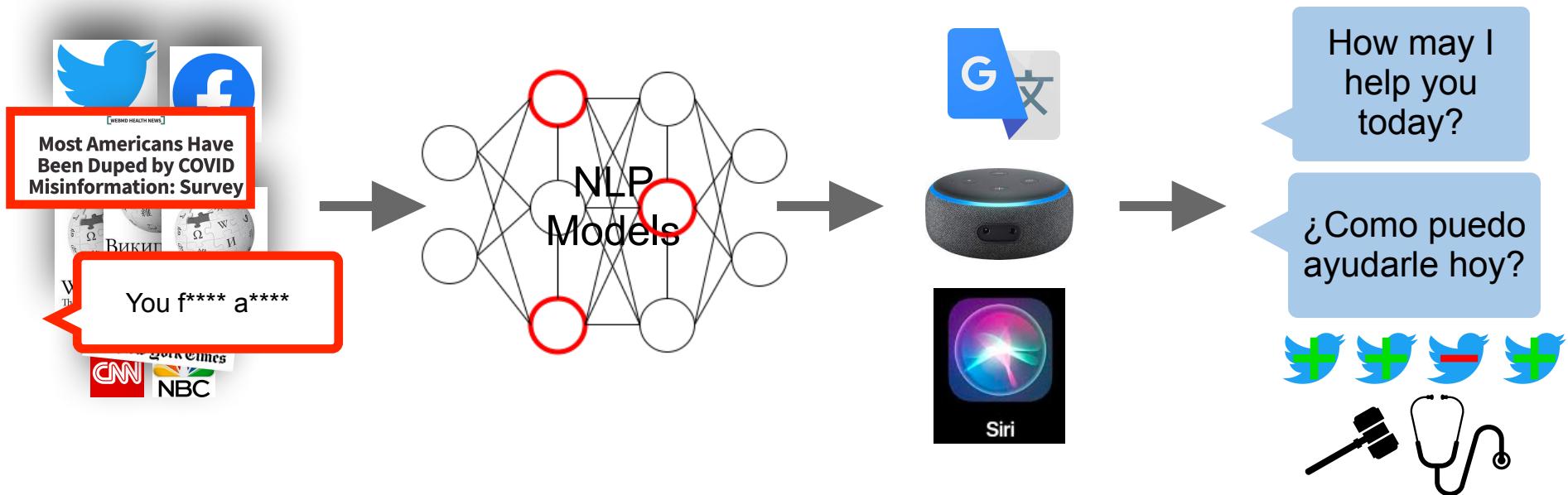
Stereotypes, and Prejudice









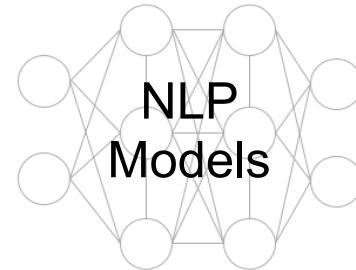
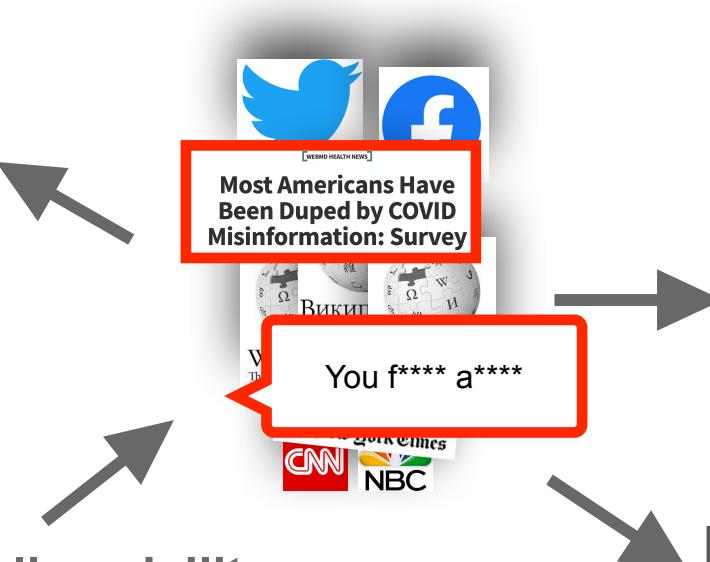


Social Science and Public Policy

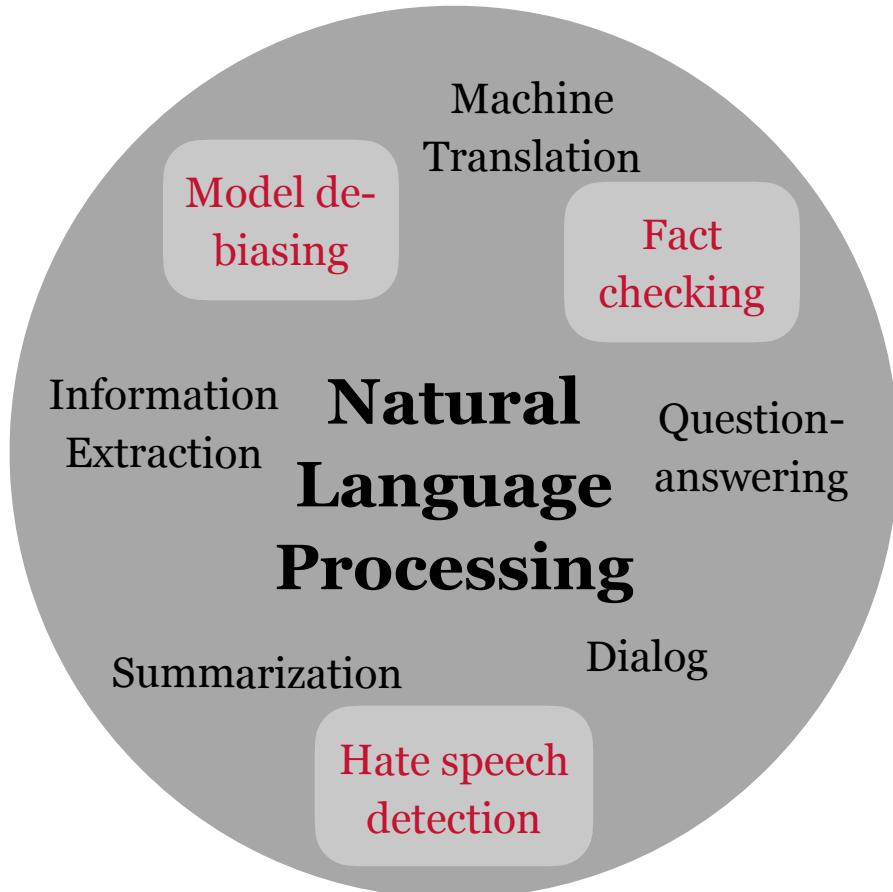
Provide insight
into human
behavior and
inform policy
decisions

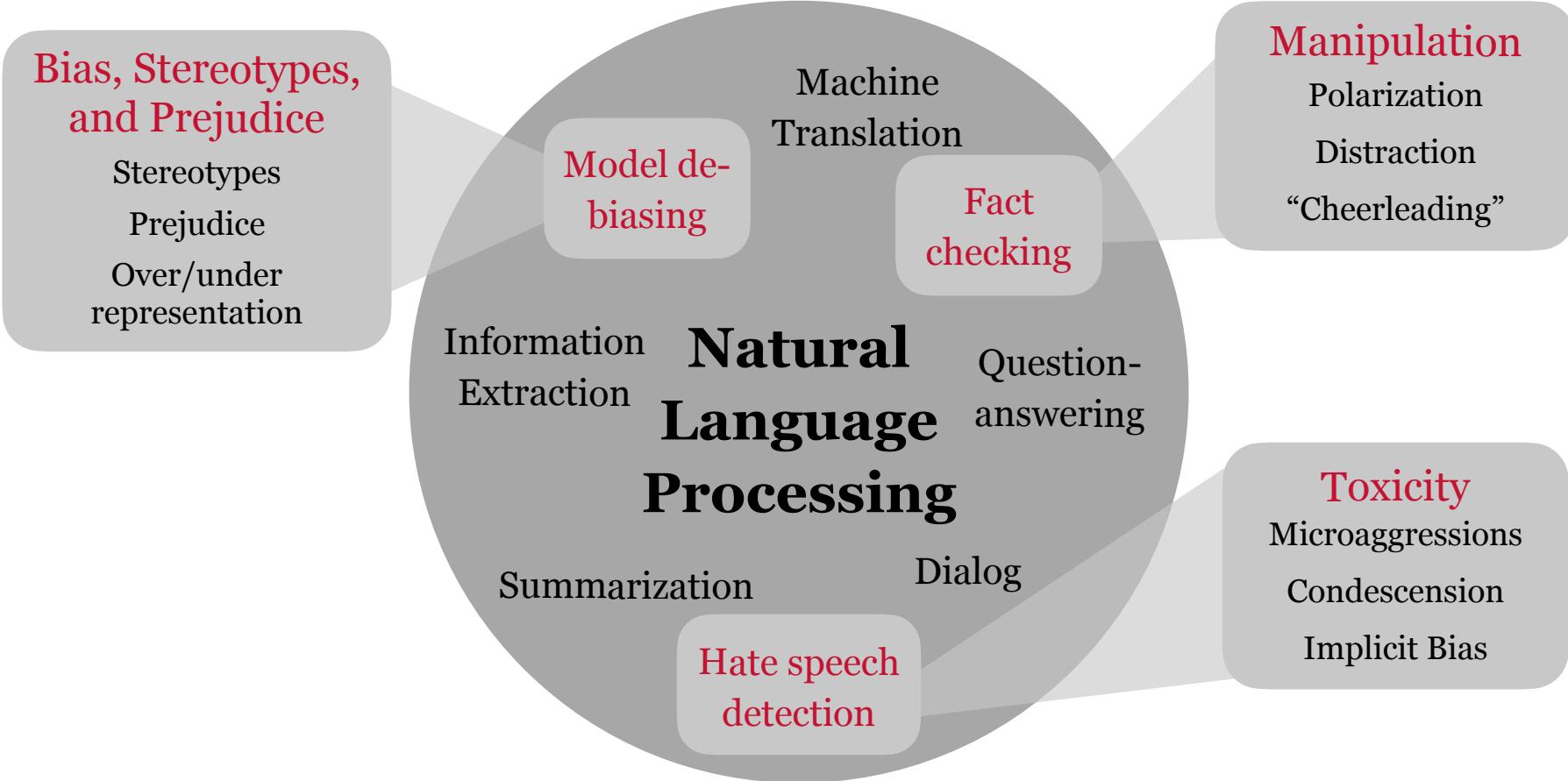


Proactive online civility
Help people avoid
accidentally creating
harmful content



**Reactive Content
Moderation**
Technology platforms,
high-stakes decision
settings





Bias, Stereotypes, and Prejudice

Stereotypes

Prejudice

Over/under representation

Information Extraction

Model debiasing

Natural Language Processing

Summarization

Hate speech detection

Machine Translation

Fact checking

Question-answering

Dialog

Manipulation
Polarization
Distraction
“Cheerleading”

Toxicity
Microaggressions
Condescension
Implicit Bias

Generalizable

Bias, Stereotypes,
and Prejudice

[ICWSM 2019](#)

ACL 2019

ICWSM 2021

ACL 2021

WWW 2022

[CRAC at EMNLP 2021](#)

FAccT 2022 (In sub.)

[PNAS 2022 \(In rev.\)](#)

Machine
Translation

Model de-
biasing

Fact
checking

Information
Extraction

Natural Language Processing

Summarization

Dialog

Hate speech
detection

Manipulation

[EMNLP 2018](#)

SocInfo 2020

Reliable

Toxicity

[EMNLP 2020](#)

[SocialNLP at ACL 2020](#)

Interpretable

Social Psychology

Generalizable

Political Science

Bias, Stereotypes,
and Prejudice

[ICWSM 2019](#)

ACL 2019

ICWSM 2021

ACL 2021

WWW 2022

[CRAC at EMNLP 2021](#)

[FAccT 2022 \(In sub.\)](#)

[PNAS 2022 \(In rev.\)](#)

Decision Science

Public Policy

Interpretable

Sociology

Model de-biasing

Information Extraction

Summarization

Hate speech detection

Natural Language Processing

Machine Translation

Fact checking

Question-answering

Dialog

Causal Inference

Economics

Manipulation

[EMNLP 2018](#)
[SocInfo 2020](#)

Reliable

Toxicity

[EMNLP 2020](#)
[SocialNLP at ACL 2020](#)

Natural Language Processing

Partnerships with industry, government, and non-profit agencies

Cross-disciplinary collaborations and education

Improve diversity in computer science research community

This talk

- Global Manipulation Strategies
 - **Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies.** Anjalie Field, Doron Klinger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. In *Proc. EMNLP'18*.
- Toxicity
 - **Unsupervised Discovery of Implicit Gender Bias**, Anjalie Field and Yulia Tsvetkov. In *Proc. EMNLP'20*.
- Future and Ongoing Work
 - Forming partnerships with industry, government, and non-profit agencies to tackle real-world problems and data

Global Manipulation Strategies: A computational analysis of propaganda



Doron Kliger



Shuly Wintner



Jennifer Pan



Dan Jurafsky



Yulia Tsvetkov

Economics
@ Haifa U

NLP
@ Haifa U

Political Science
@ Stanford

CSS, NLP
@ Stanford

NLP
@ UW

- **Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies.** Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. In *Proc. EMNLP'18*.





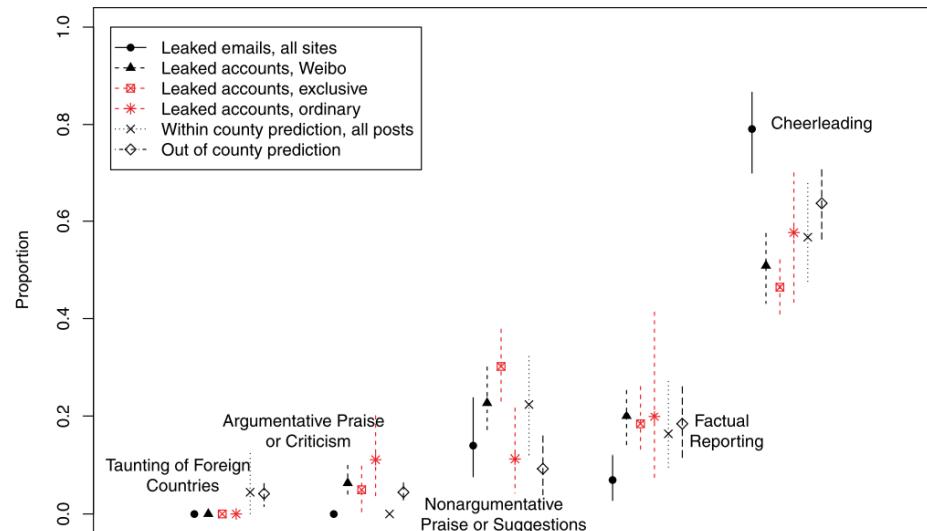
Targeted manipulation of elections and foreign politics

How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument

GARY KING *Harvard University*

JENNIFER PAN *Stanford University*

MARGARET E. ROBERTS *University of California, San Diego*



Flooding social media with positive messages to deter collective action

The Surprising Nuance Behind the Russian Troll Strategy

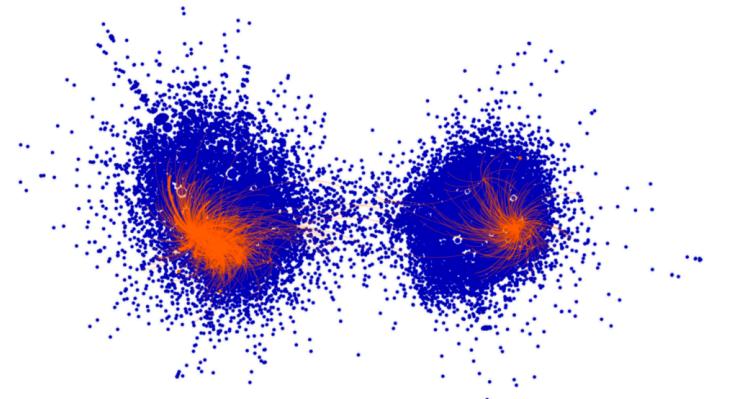
We set out to study internet discourse around #BlackLivesMatter — instead, we were unintentionally learning about the Russian information operation to undermine democracy



Kate Starbird Oct 20, 2018 · 10 min read *



Promoting polarizing content to de-stabilize regimes and for political gain



Left-leaning

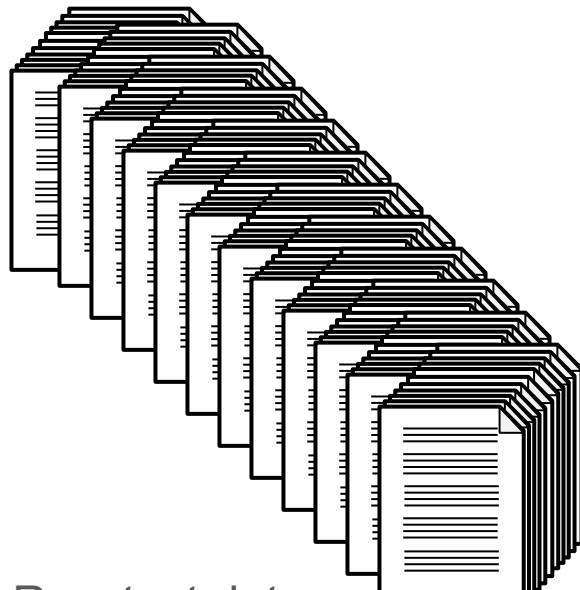
Right-leaning



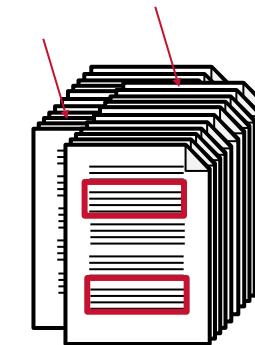
Left-leaning

Right-leaning

How can we detect this type of media manipulation at scale?



Raw text data
in different languages



Documents, topics, and phrases
that contain **manipulative content**

Theories from communications research

- Agenda setting
 - *What* topics are covered
- Framing
 - *How* topics are covered
- Priming
 - What *effects* the reporting has on public opinion

Journal of Communication ISSN 0021-9916

ORIGINAL ARTICLE

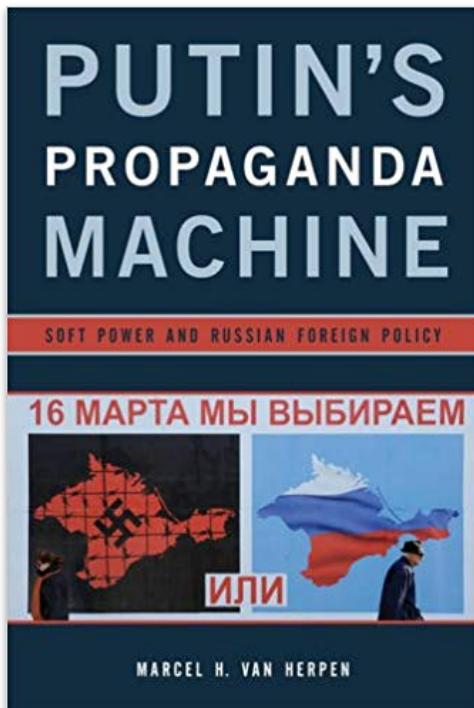
Framing Bias: Media in the Distribution of Power

Robert M. Entman

School of Media and Public Affairs, The George Washington University, Washington, DC 20052

“agenda setting, framing and priming fit together as tools of power”

Investigation of Russian news



The screenshot shows a news page from 'ИЗВЕСТИЯ' (Izvestiya). The top navigation bar includes categories like 'ПОЛИТИКА' (Politics) and 'ОБЩЕСТВО' (Society). The main headline reads: 'Появились кадры с места обрушения строящегося забора в Подмосковье' (Footage has appeared from the site of the collapse of a building under construction in the Moscow region). Below it, another article is visible with the headline: 'МИД РФ показал примеры дезинформации кампании западных' (The Russian Foreign Ministry showed examples of disinformation from Western media). A sidebar on the right contains the text: 'Прокуратура начала проверку по факту обрушения бетонного забора в Подмосковье' (The procuratorate has started an investigation into the collapse of a concrete fence in the Moscow region).

Agenda setting

“

...the media may not be successful much of the time in telling people *what to think*, but is stunningly successful in telling its readers *what to think about*”

(Cohen, 1963)

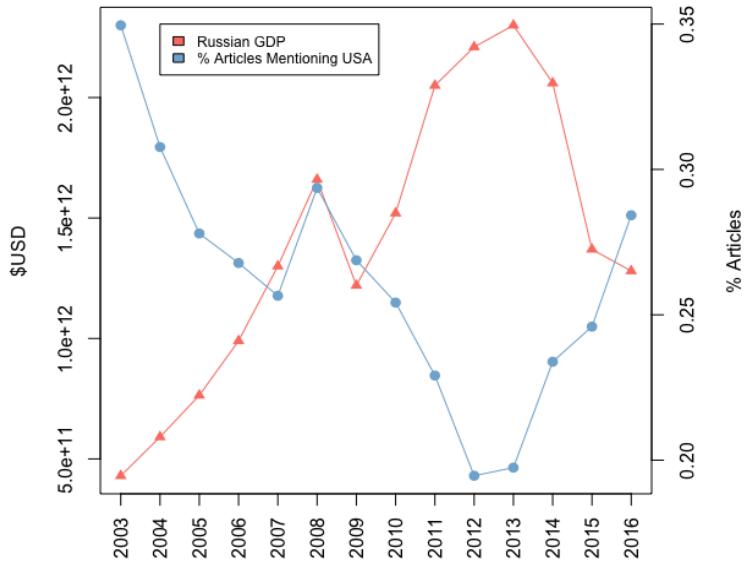
When might (government-influenced) media want to shape what people think about?

Benchmark against economic indicators

- Can hypothesize that we will see more manipulation strategies when the country is “doing poorly”
 - Government wants to distract public or deflect blame
- Measure of “doing poorly”
 - State of the economy (GDP and stock market)

Agenda Setting: Do Russian news articles discuss foreign countries (the U.S.) more during economic downturns?

Frequency of mentions of the U.S.



Pearson's correlation
with articles that
mention the U.S.

RTSI (Monthly, rubles)

-0.54

GDP (Quarterly, USD)

-0.69

GDP (Yearly, USD)

-0.83

Granger causality

$$C(w_t) = \sum_{i=1}^m \alpha_i(C(w_{t-i})) + \sum_{j=1}^n \beta_j(C(r_{t-j}))$$

C() percent change

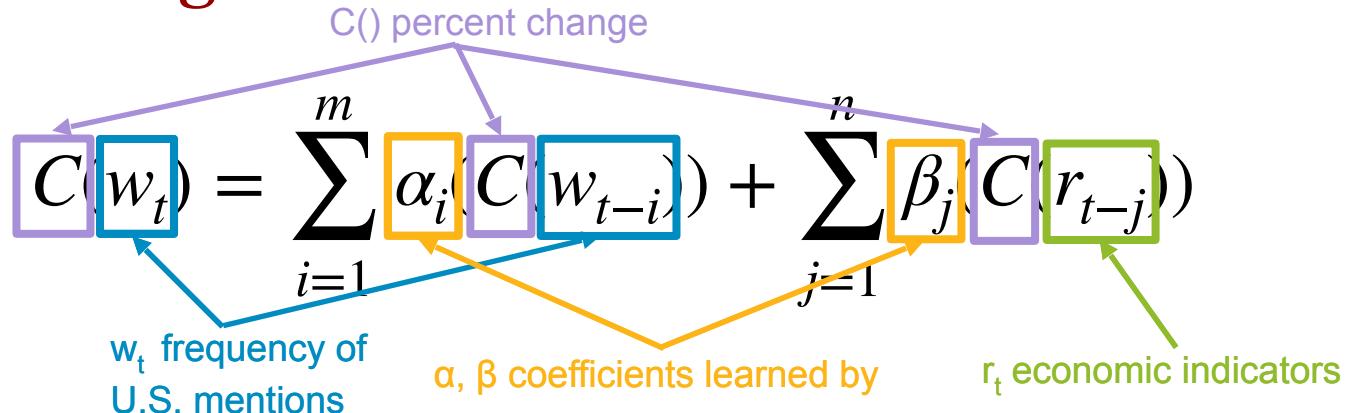
w_t frequency of U.S. mentions

α, β coefficients learned by

r_t economic indicators

The diagram illustrates the Granger causality equation. At the top, a purple bracket labeled "C() percent change" spans the entire right side of the equation. Below it, the term $C(w_t)$ is enclosed in a purple box. To its left, a blue arrow points from the text "w_t frequency of U.S. mentions" to the w_t term. Further left, another blue arrow points from the text "C() percent change" to the $C(w_t)$ term. The first summation term, $\sum_{i=1}^m \alpha_i(C(w_{t-i}))$, is enclosed in a yellow box. A blue arrow points from the text "w_t frequency of U.S. mentions" to the w_{t-i} term inside this box. An orange arrow points from the text " α, β coefficients learned by" to the α_i coefficient. The second summation term, $\sum_{j=1}^n \beta_j(C(r_{t-j}))$, is enclosed in a yellow box. An orange arrow points from the text " α, β coefficients learned by" to the β_j coefficient. This box is also connected by a green arrow to the term $C(r_{t-j})$, which is enclosed in a green box. Finally, a green arrow points from the text "r_t economic indicators" to the r_{t-j} term.

Agenda setting evidence

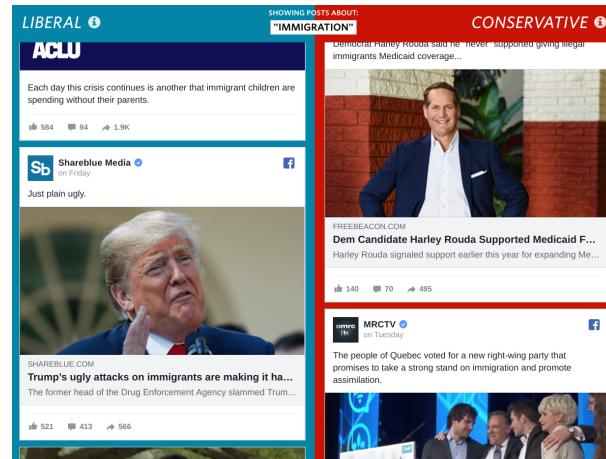


	α;β	p-value
w _{t-1}	-0.320	0.00005
w _{t-2}	-0.301	0.0001
r _{t-1}	-0.369	0.024
r _{t-2}	-0.122	0.458

Framing

“To frame is to *select some aspects of a perceived reality and make them more salient*”, e.g. to “promote a particular...interpretation” (Entman, 1993)

- Topic level
 - Abortion is a moral issue
 - Abortion is a health issue
- Word level
 - “Pro-life” vs “pro-choice”



Infer Russian media frames using distant

- **Media Frames Corpus** (Boydston et al. 2014; Card et al. 2015)
 - ~ 11,000 articles annotated with 14 policy-oriented frames

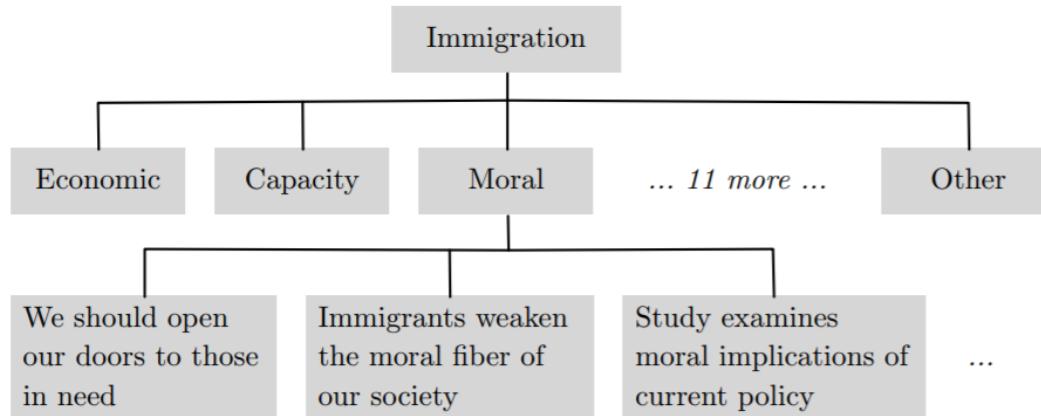
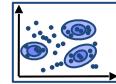


Figure 1: Illustration of hierarchical policy frames coding scheme: Immigration.

How can we adapt English framing annotations to Russian news articles?

Annotation of *Izvestia* articles with MFC



**Extract lexicons
from MFC
(PMI Scores)**

→
**Translate lexicons
into target language**

→
**Query-expansion
(word embeddings)
to adapt lexicons to
target corpus**

→
**Identify document-
level frames**

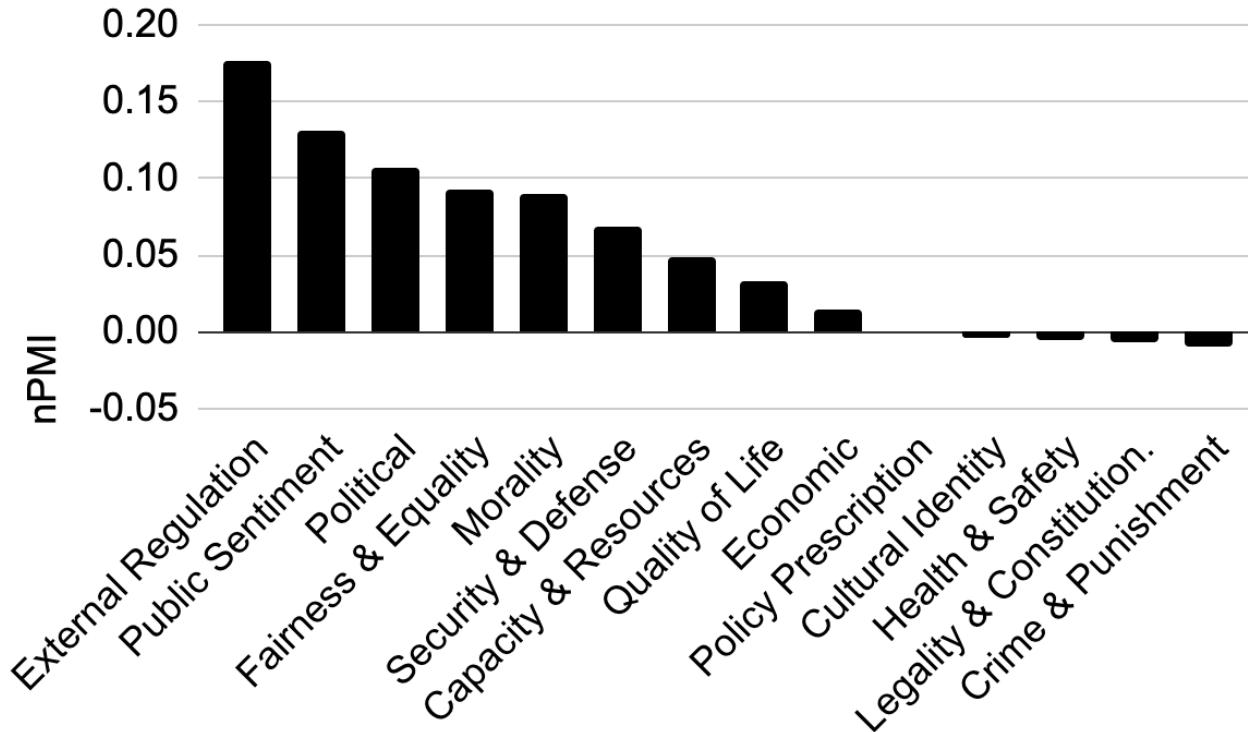
“Dollars”

“долларов”

“рублей”
(rubles)

Economic

Analysis: which frames are most salient in U.S. focused articles?



Analysis of specific frames

- Which frames and words become more salient after downturns and less salient after upturns?
 - Security and Defense: bombs, missiles, Guantanamo, North Korea, Iraq
 - What types of statements are said about the U.S.?
 - “Nazi vultures... bombing of the [redacted] [villainizing the U.S.] city. The barbaric over the world”
 - “The U.S. [redacted] [describing threats to the U.S.] y to hide its crimes, e [redacted] [describing threats to the U.S.]
 - “The U.S. mission in Guantanamo outside of [redacted] [describing threats to the U.S.]”
- [promoting the Russian military over the U.S. military]

Summary: a computational analysis of propaganda

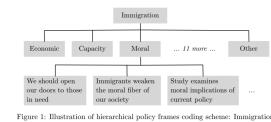
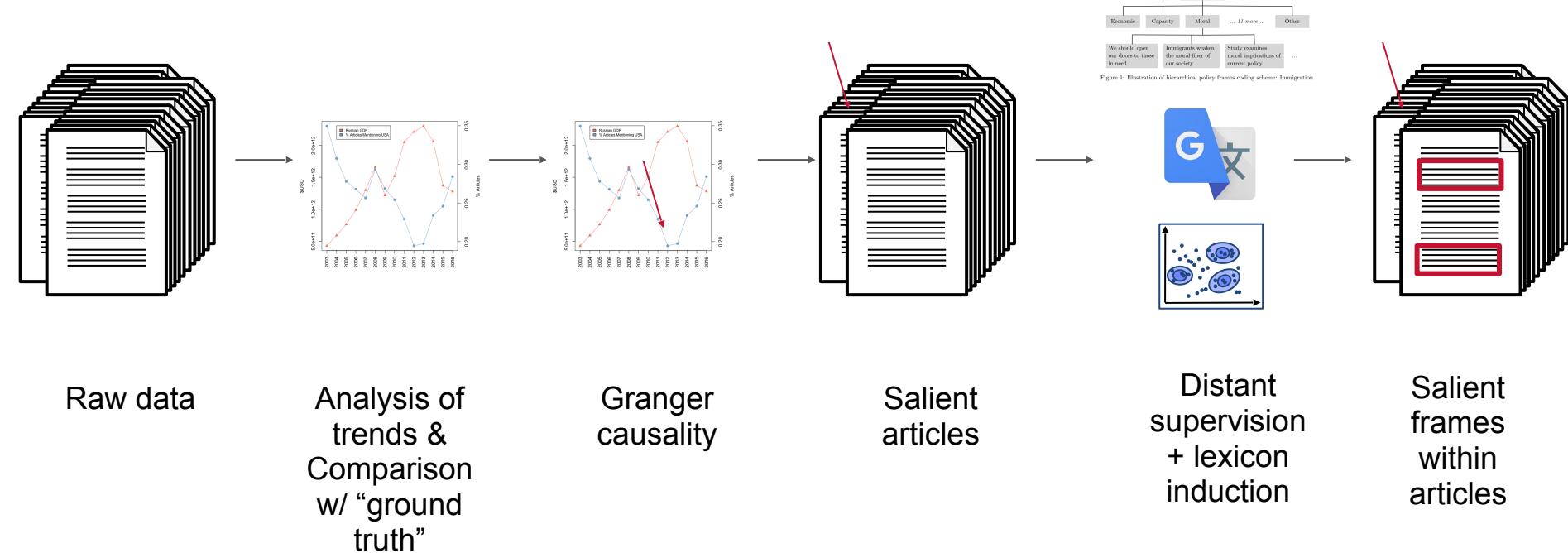


Figure 1: Illustration of hierarchical policy frames coding scheme: Immigration.

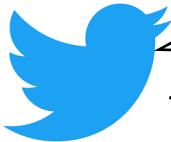
Outcomes and Impacts

- What are societal benefits from this work?
 - Publicizing propaganda strategies reduces credibility of unreliable sources (Roberts, 2020)
 - Facilitates political science research that can inform public policy
- What are NLP contributions to this work?
 - Characterizing harms in text informs NLP ethics — *without it, we don't know what we're looking for*
 - NLP tasks and methodology for identifying subtle connotations
 - Follow-up work on detecting and analyzing agenda-setting and framing

'Fiction is outperforming reality': how YouTube's algorithm distorts truth

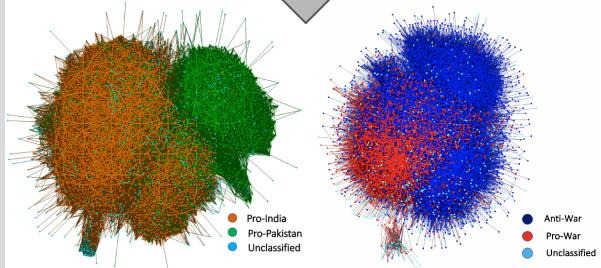
An ex-YouTube insider reveals how its recommendation algorithm promotes divisive clips and conspiracy videos. Did they harm Hillary Clinton's bid for the presidency?

● [The methodology behind this story](#)
by [Paul Lewis](#) in San Francisco

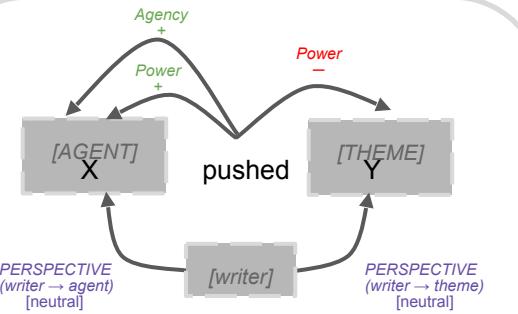


हमारे बच्चों के लिए
शांति, हमारे भविष्य के
लिए शांति !!

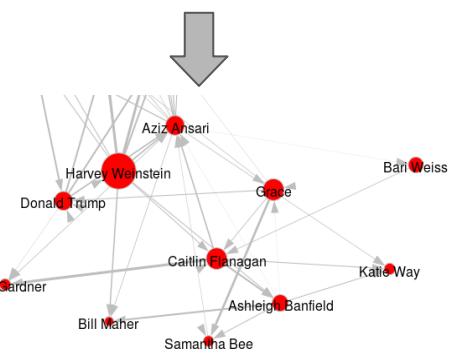
ہم بے صبری سے انتظار کر
ربے بیس
آپ کی انتقامی کا روانی



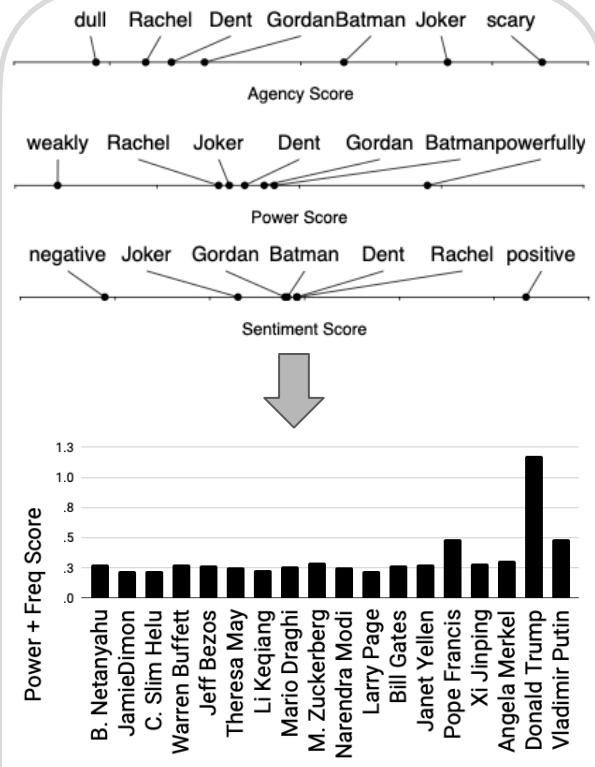
Tyagi*, Field*, et al. (2020) **A Computational Analysis of Polarization on Indian and Pakistani Social Media.** *SocInfo*
[Best Paper Nominated]



Connotation Frames



Field et al. (2019) Contextual
Affective Analysis: **A Case Study**
of People Portrayals in Online
#MeToo Stories. *ICWSM*



Field and Tsvetkov. (2019) **Entity-Centric Contextual Affective Analysis.** *ACL*

This Talk

- Global Manipulation Strategies
 - **Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies.** Anjalie Field, Doron Klinger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. In *Proc. EMNLP'18*.
- Toxicity
 - **Unsupervised Discovery of Implicit Gender Bias**, Anjalie Field and Yulia Tsvetkov. In *Proc. EMNLP'20*.
- *Future and Ongoing Work*
 - *Forming partnerships with industry, government, and non-profit agencies to tackle real-world problems and data*

Goal: Identify text containing (subtle)



[Original Writer]

November 12, 2021 ·

...

Bob and I join Bill Hemmer on America's Newsroom to discuss whether or not...



[Commenter]

I like Bob, but you're hot, so kick his butt

Like · Reply · 9w



Alexandria Ocasio-Cortez

December 25, 2021 at 10:33 AM ·

...

Merry Christmas and happy holidays to NY-14 and beyond! Wishing you and yours a safe and healthy holiday season and a wonderful New Year.



How about you adopt some unfortunate kids ? That would actually help & be un - selfish / un self serving, & help the unfortunate, I'll be really awaiting your reply , thanks for your attention ❤

...



Yes , you could care yourself. You want al. , A shame your father di blessing not to have yo

Like · Reply · 2w



Like · Reply · 3w

~~your soy soy soy with your mom trying to teach you something?? Dreaming of something for yourself?? Bet you struck out though because Republican men DON'T want to do ANYTHING WITH YOU!~~

...

something?? Dreaming of something for yourself?? Bet you struck out though because Republican men DON'T want to do ANYTHING WITH YOU!



Like · Reply · 2w

Resume

CONFERENCE & JOURNAL PUBLICATIONS

- Anjali Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. "A Survey of Race, Racism, and Anti-Racism in NLP" (2021). Annual Meeting of the Association for Computational Linguistics (ACL), <https://aclanthology.org/2021.acl-long.149.pdf>
- Chas Young Park*, Xiru Yan*, Anjali Field*, and Yulia Tsvetkov. "Multilingual Contextual Affective Analysis of LGBT People Portrayals in Wikipedia" (2021). International AAAI Conference on Web and Social Media (ICWSM), <https://arxiv.org/abs/2010.10820>
- Anjali Field and Yulia Tsvetkov. "Unsupervised Discovery of Implicit Gender Bias" (2020). Conference on Empirical Methods in Natural Language Processing (EMNLP), <https://aclanthology.org/2020.emnlp-main.44/>
- Aman Tyagi*, Anjali Field*, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M. Carley. "A Computational Analysis of Polarization on Indian and Pakistani Social Media" (2020). International Conference on Social Informatics (SocInfo) (nominated for Best Paper), <https://arxiv.org/abs/2005.09803>
- Anjali Field and Yulia Tsvetkov. "Entity-Centric Contextual Affective Analysis" (2019). Annual Meeting of the Association for Computational Linguistics (ACL), <https://www.acweb.org/anthology/P19-1243.pdf>
- Anjali Field, Gayatri Bhagat, Yulia Tsvetkov. "Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories" (2019). International AAAI Conference on Web and Social Media (ICWSM), <https://www.aaai.org/ojs/index.php/ICWSM/article/view/3358/2226>.

WORKSHOP PUBLICATIONS

- Nupoor Gandhi, Anjali Field, and Yulia Tsvetkov. "Improving Span Representation for Domain-adapted Conference Resolution" (2021). CRAC at EMNLP <https://arxiv.org/pdf/2109.09811.pdf>
- Mengzhou Xia, Anjali Field, and Yulia Tsvetkov. "Demoting Racial Bias in Hate Speech Detection" (2020). SocialNLP at ACL <https://aclanthology.org/2020.socialnlp-1.2/>

INVITED TALKS

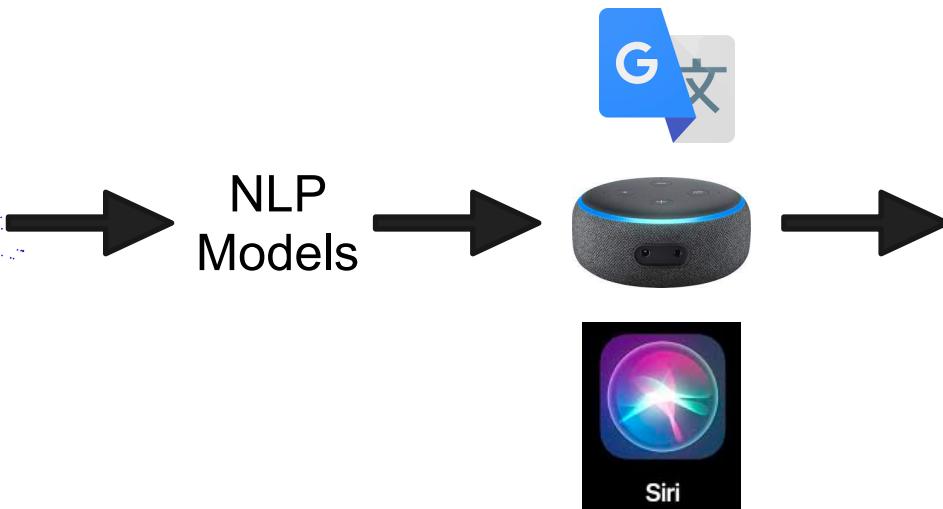
- NLP Methods for Identifying Gender Bias 2021
Stanford Women in CS
Detection of Stereotypes, Bias, and Prejudice in Text 2021
Stanford NLP Seminar
Reducing Confounding Variables in Social Text Processing 2021
Educational Testing Service (ETS)
Unsupervised Discovery of Implicit Gender Bias 2021
PhD Introductory Meeting at University of Washington

TEACHING

- Guest lecture for Undergraduate Seminar in Ethics and Fairness in AI Spring 2021
■ University of Pittsburgh, "Contextual Affective Analysis"
TA for Algorithms for NLP Fall 2019
■ Carnegie Mellon University, Facilitated homework assignments on topics like language modeling; delivered lectures and recitations
TA for Computational Ethics for NLP (11-830) Spring 2019
■ Carnegie Mellon University, Facilitated homework assignments on topics like hate speech detection; delivered lectures on propaganda and bias; advised projects on fake news and media bias
Guest lecture for Algorithms for NLP Fall 2018
■ Carnegie Mellon University, "Computational Social Science"

She's qualified but she
seems **really aggressive**

I like her ideas but she
wasn't very friendly.
Would it have killed her
to smile?



“Oh, you work
at an office? I
bet you’re a
secretary”

“Total tangent I
know, but you’re
gorgeous”

Need to develop new models

Our goal: detect subtle gender biases like microaggressions, objectifications, and condescension in 2nd-person text

- “Oh, you work at an office? I bet you’re a secretary”
- “Total tangent I know, but you’re gorgeous”

Current classifiers that detect hate speech, offensive language, or negative sentiment cannot detect these comments

Naive Approach: Supervised Classification



I like Bob, but you're hot,
so kick his butt

Like · Reply ·



Thanks so much **Ma'am!**

Like · Reply ·



I'd vote for you if I lived in
Massachusetts

Like · Reply ·



...a good way to celebrate
Title IX, too!

Like · Reply ·



Naive Approach: Supervised Classification



I like Bob, but you're hot,
so kick his butt

Like · Reply ·



Thanks so much **Ma'am!**

Like · Reply ·



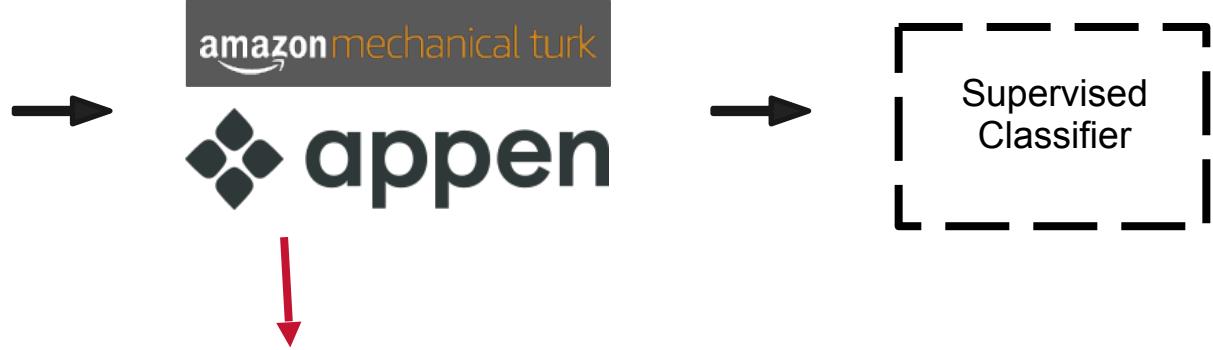
I'd vote for you if I lived in
Massachusetts

Like · Reply ·



...a good way to celebrate
Title IX, too!

Like · Reply ·



Problem: Biases are subtle, implicit, and context-dependent

Proposed approach: Comments contain gender bias if they are highly predictive of gender

Would the addressee have received different text if their gender were different?

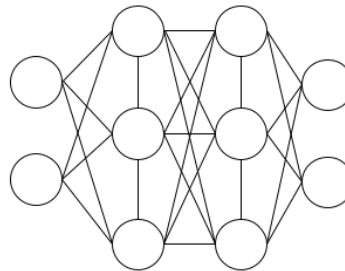
Proposed approach: Comments contain gender bias if they are highly predictive of gender

- Train a classifier that predicts the **gender** of the **person the text is addressed to**
- If the classifier makes a prediction with high confidence, the text likely contains bias



I like Bob, but you're hot,
so kick his butt

Like · Reply ·



Addressed to **Man**



Addressed to **Woman**

If a comment is very likely to be addressed to a woman, and is very unlikely to be addressed to a man, it probably contains gender bias.

Challenge: Text main contain *confounds* that are predictive of gender, but not indicative of gender bias



I like Bob, but you're hot,
so kick his butt

Like · Reply ·



Thanks so much **Ma'am!**

Like · Reply ·



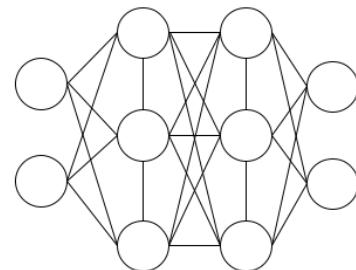
I'd vote for you if I lived in
Massachusetts

Like · Reply ·



...a good way to celebrate
Title IX, too!

Like · Reply ·



→ Addressed to **Woman**

→ Addressed to **Woman**

→ Addressed to **Woman**

→ Addressed to **Woman**

Challenge: Text main contain *conounds* that are predictive of gender, but not indicative of gender bias

- **Overtly gendered words**
- **Preceding context in the conversation**
- **Traits of people (other than gender) in the conversation**



Saturday is the 40th anniversary of **Title IX**...

Like · Reply ·



...a good way to celebrate
Title IX, too!

Like · Reply ·



I'd vote for you if I lived in
Massachusetts

Like · Reply ·



Bob and I join Bill Hemmer on
America's Newsroom to discuss whether
or not...

Like · Reply ·



I like Bob, but you're hot, so
kick his butt

Like · Reply ·



Thanks so much Ma'am!

Like · Reply ·

Proposed Model: Comments contain bias if they are highly predictive of gender *despite confound control*

- Substitute overt indicators: replace overtly gendered terms with neutral ones



Preceding context is an *observed* confounding variables

Writer_Gender: F



Saturday is the 40th anniversary of **Title IX**! I'm celebrating with a Sat morning run - ladies please respond below if you want to join

Like · Reply ·



Wish I could ! Already have plans for a bike ride and breakfast with some awesome ladies - a good way to celebrate **Title IX**, too!

Like · Reply ·



Would love to!

Like · Reply ·

Writer_Gender: M



Any deal with **Iran** — a nation that the United States cut off diplomatic ties with 35 years ago — must protect America's interests at home

Like · Reply ·



Iran might be a free, democratic nation today, if not for decades of American interference.

Like · Reply ·

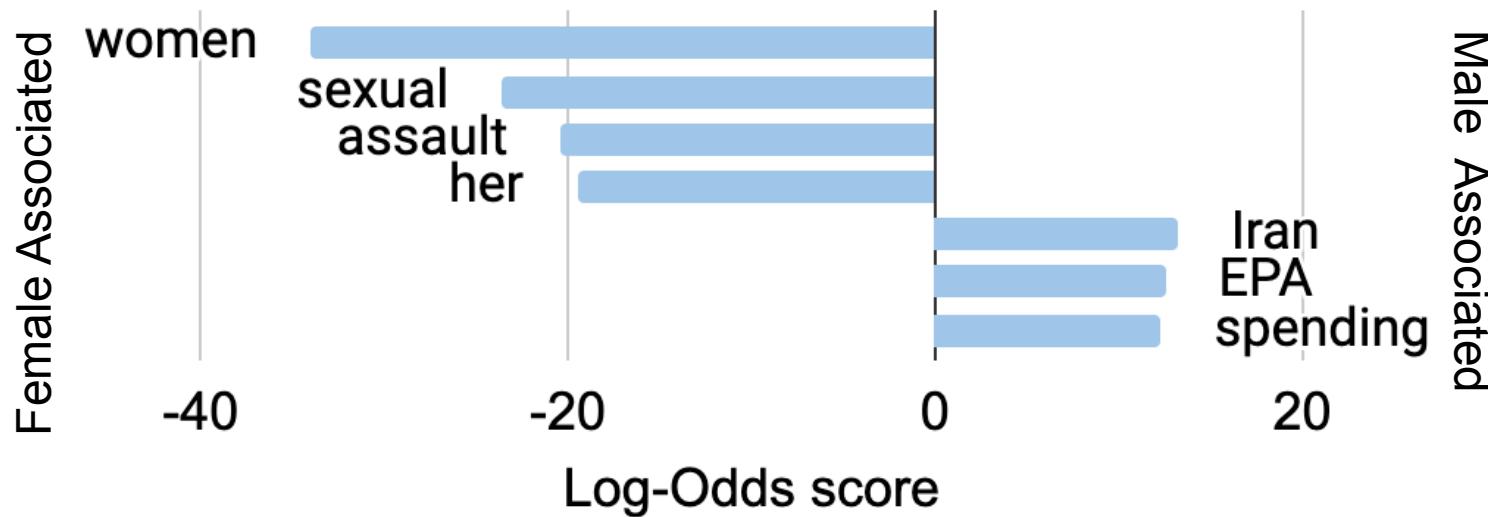


That's for sure! Worst deal he could make! We can't trust **Iran** and America knows it !!!!!

Like · Reply ·

Key problem: Men and women post different content, which is reflected in their replies

Preceding context is an *observed* confounding variables



Propensity matching for *observed* confounding variables

Writer_Gender: F

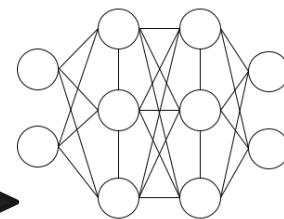
Saturday is the 40th anniversary of **Title IX**! I'm celebrating with a Sat morning run - ladies please respond below if you want to join.

Writer_Gender: M

Any deal with **Iran** — a nation that the United States cut off diplomatic ties with 35 years ago — must protect America's interests at home and abroad.

Writer_Gender: F

My overriding concern is whether or not the agreement is in the national security interest of the United States. **Iran** must be blocked from proceeding any further towards



Text classifier to predict WRITER_GENDER



$e_i = P(W.\text{Gender}_i = F | Post_i) \approx 0.91$

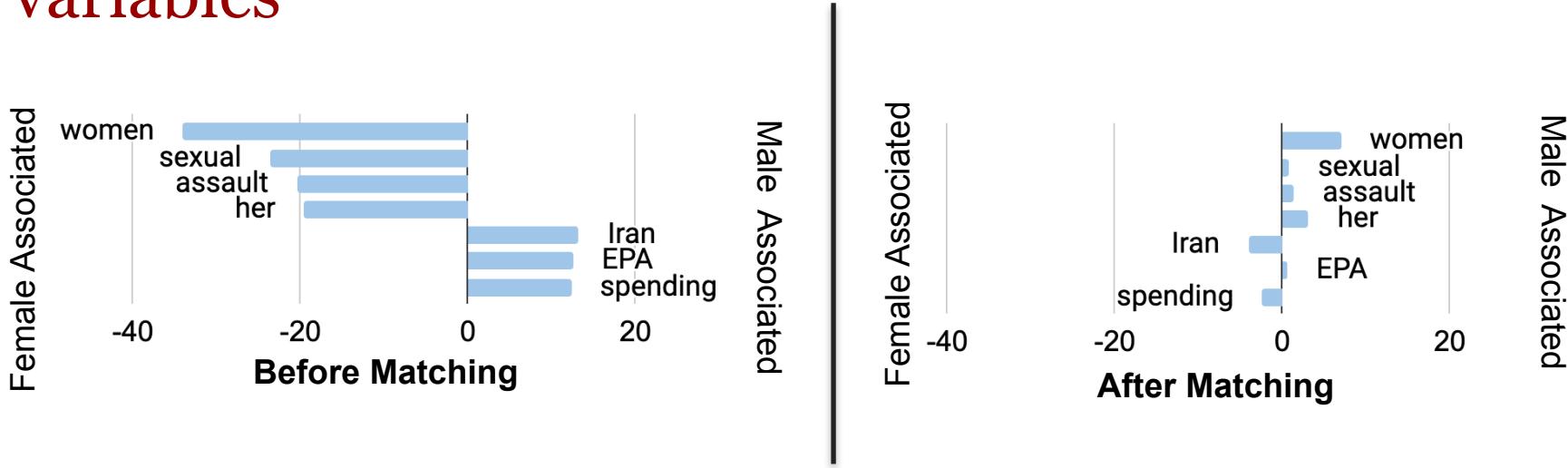
$$|e_i - e_l| \geq c \forall l$$

$e_j = P(W.\text{Gender}_j = F | Post_j) \approx 0.33$

$e_k = P(W.\text{Gender}_k = F | Post_k) \approx 0.32$

$$\operatorname{argmin}_j |e_k - e_j|$$

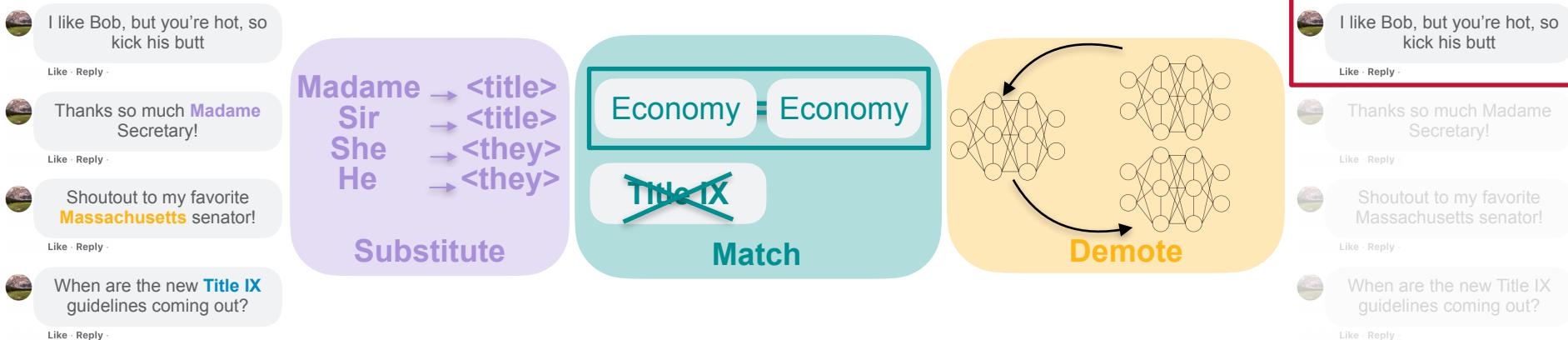
Propensity matching for *observed* confounding variables



Propensity matching breaks associations between gender and context in the training data

Proposed Model: Comments contain bias if they are highly predictive of gender *despite confound control*

- Substitute overt indicators
- Balance observed confounders through propensity matching
- Demote latent confounders through adversarial training



Adversarial training for *latent* confounding variables

- Comments may references traits of the addressee (such as occupation, nationality, nicknames, etc.) that are correlated with gender
- Difficult to enumerate all of them
- Often unique to individuals (difficult to make matches)



A vote for **Liz** Warren is a vote for a saner **Massachusetts** and a saner America.

[Like](#) · [Reply](#) ·



Go **Lizzie** go!!!! Good luck next Tuesday. **Massachusetts** will be lucky to have you as their Senator.

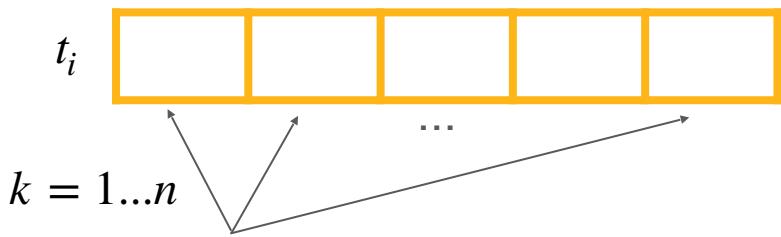
[Like](#) · [Reply](#) ·



'**Lizbeth**... I'd vote for you if I lived in **Massachusetts**, in a heartbeat

[Like](#) · [Reply](#) ·

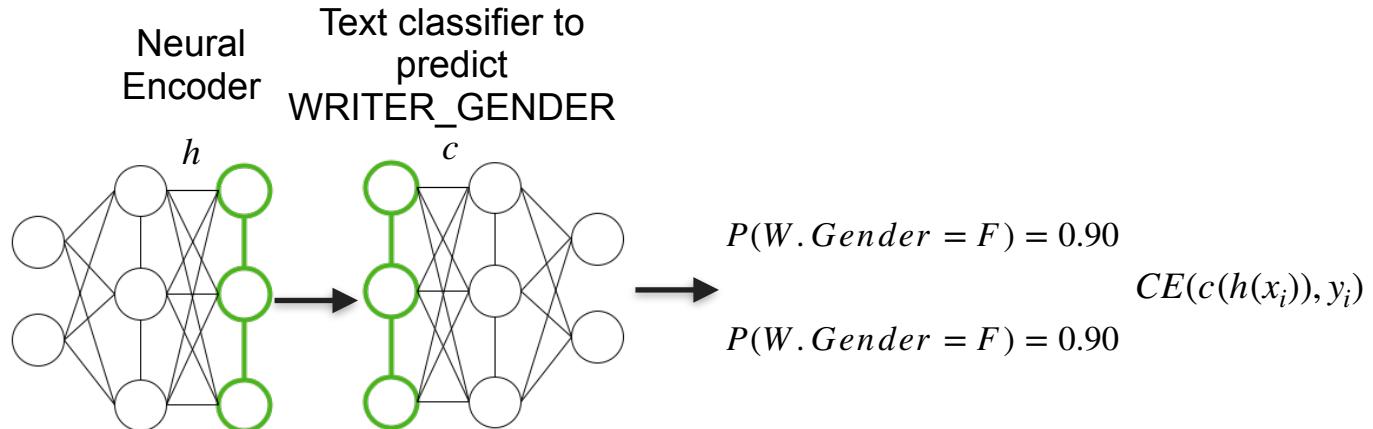
Represent latent confounding variables as a vector



$$\begin{aligned} p(\text{addressee} = k | \text{comment}) &\propto p(\text{addressee} = k) p(\text{comment} | \text{addressee}) \\ &= p(\text{addressee} = k) \prod_{w_i \in \text{comment}} p(w_i | k) \end{aligned}$$

Adversarial training for *latent* confounding variables

-  I like Bob, but you're hot,
so kick his butt
Like · Reply ·
-  Shoutout to my
favorite **Massachusetts**
senator!
Like · Reply ·



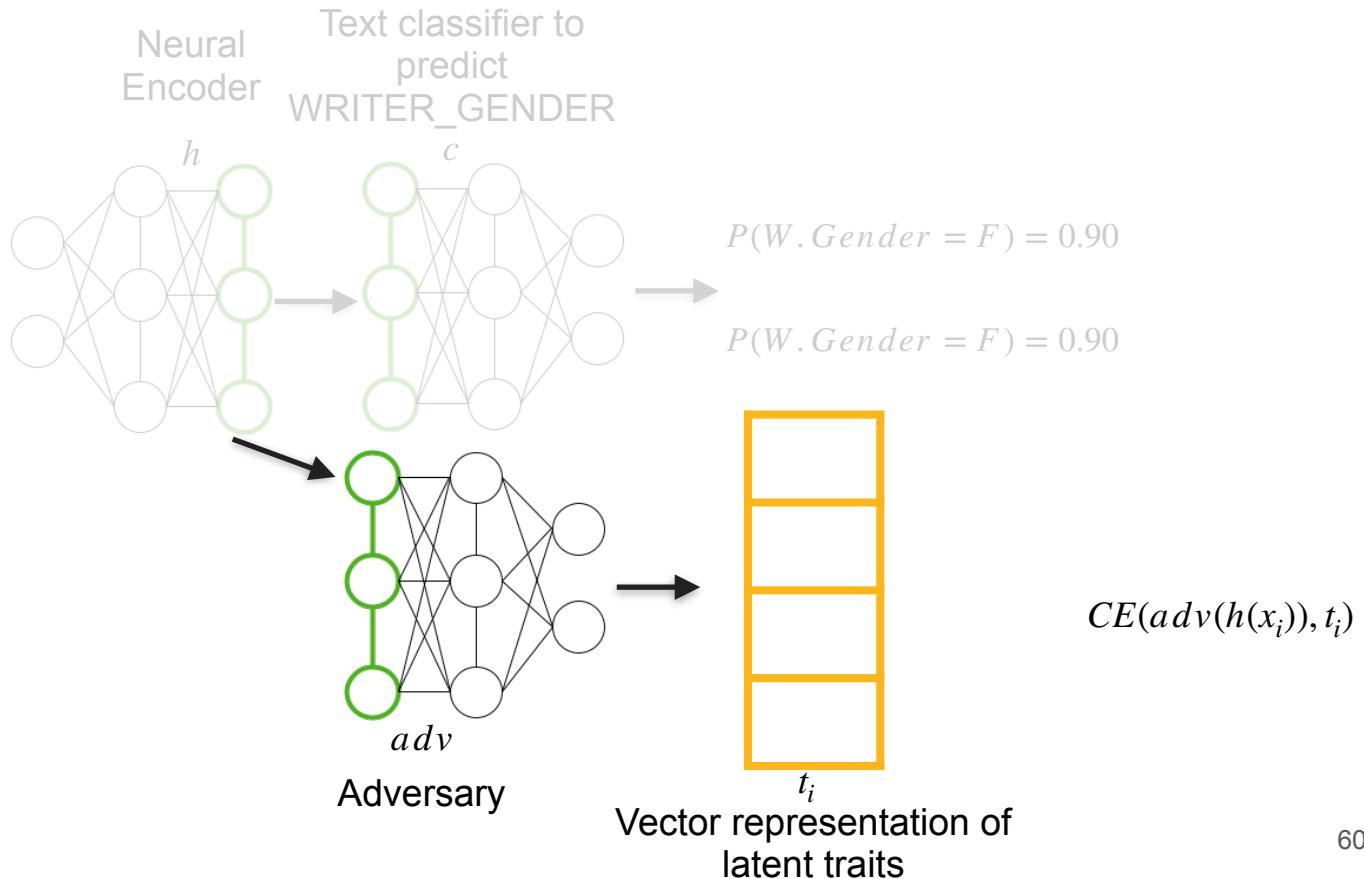
Adversarial training for *latent* confounding variables

 I like Bob, but you're hot,
 so kick his butt

Like · Reply

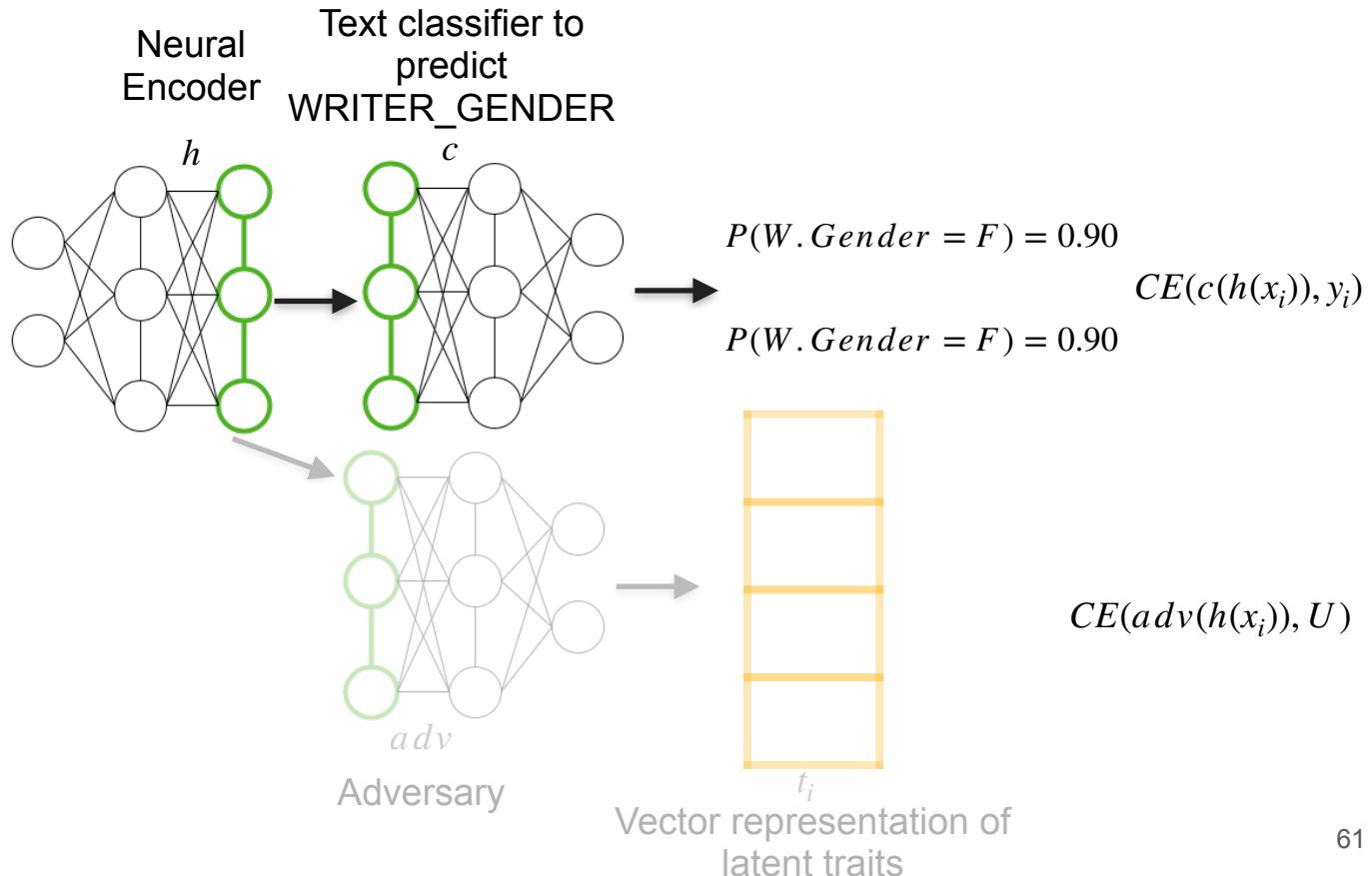
 Shoutout to my
favorite **Massachusetts**
senator!

Like · Reply



Adversarial training for *latent* confounding variables

-  I like Bob, but you're hot,
so kick his butt
Like · Reply ·
-  Shoutout to my
favorite **Massachusetts**
senator!
Like · Reply ·

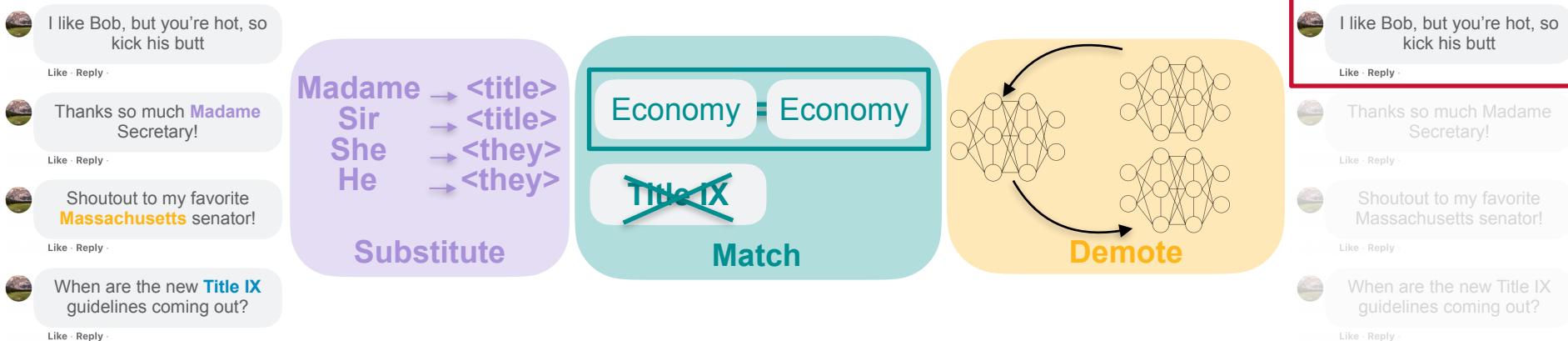


Evaluation: Performance improvement on held-out data

	Public Figures		Politicians	
	F1	Acc.	F1	Acc.
base	74.9	63.8	23.2	73.2
+demotion	76.1	65.1	17.4	77.1
+match	65.4	56.0	28.5	46.7
+match+demotion	68.2	59.7	28.8	51.4

Proposed Model: Comments contain bias if they are highly predictive of gender *despite confound control*

- Substitute overt indicators
- Balance observed confounders through propensity matching
- Demote latent confounders through adversarial training



Findings: characteristics of bias against women politicians

Influential words:

- Competence and domesticity
- ‘Force’, ‘situation’, ‘spouse’, ‘family’, ‘love’

Examples:

- “DINO I hope another real Democrat challenges you next election”
- “I did not vote for you and have no clue why anyone should have. You do not belong in politics”

Findings: characteristics of bias against women

Influential words:

- Appearance and sexualization
- ‘beautiful’, ‘love’, ‘sexo’

Examples:

- “Total tangent I know but, you’re gorgeous.”
- “I like Bob, but you’re hot, so kick his butt.”

Outcomes and impact

- Follow-up work investigating impacts of microaggressions in training data on NLP systems
- Funding (280K) from  to identify and mitigate implicit bias in workplace communications
- Ongoing funded (150K + 60K) collaboration with  Allegheny County Department of Human Services on identifying implicit biases in child welfare cases

This Talk

- Global Manipulation Strategies
 - **Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies.** Anjalie Field, Doron Klinger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. In *Proc. EMNLP'18*.
- Toxicity
 - **Unsupervised Discovery of Implicit Gender Bias**, Anjalie Field and Yulia Tsvetkov. In *Proc. EMNLP'20*.
- **Future and Ongoing Work**
 - *Forming partnerships with industry, government, and non-profit agencies to tackle real-world problems and data*

Social Psychology

Generalizable

Political Science

Sociology

Economics

Bias, Stereotypes,
and Prejudice

Stereotypes

Prejudice

Over/under
representation

Decision Science

Public Policy

Interpretable

Model de-
biasing

Information
Extraction

Summarization

Hate speech
detection

Dialog

Natural Language Processing

Fact
checking

Question-
answering

Manipulation
Polarization
Distraction
“Cheerleading”

Reliable

Toxicity
Microaggressions
Condescension
Implicit Bias

Causal Inference

Natural Language Processing

Partnerships with industry, government, and non-profit agencies

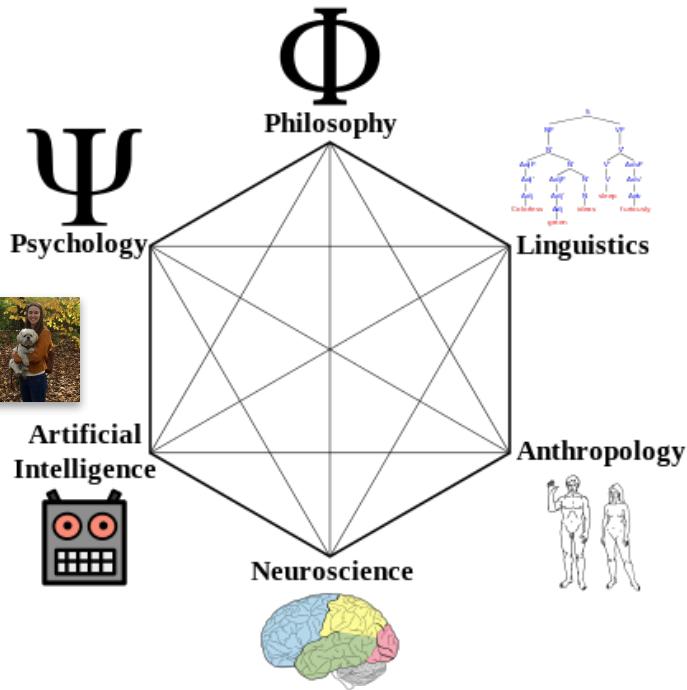
Cross-disciplinary collaborations and education

Improving diversity in CS research community

Cross-disciplinary **collaborations** and education



- Machine Learning
- Statistics
- Linguistics
- Public policy
- Social psychology
- Political science
- Economics
- Philosophy
- ...



Cross-disciplinary collaborations and **education**

Computational Ethics for NLP

CMU CS 11830, Spring 2020

T/Th 10:30-11:50am, SH 214

Yulia Tsvetkov (office hours by appointment), ytsvetko@cs.cmu.edu

Alan W Black (office hours: Wednesdays 12-1pm, Zoom link on [Piazza](#)), awb@cs.cmu.edu

TA: **Anjalie Field** (office hours by appointment), anjalief@cs.cmu.edu

TA: **Michael Miller Yoder** (office hours by appointment), yoder@cs.cmu.edu

[Summary](#) [Announcements](#) [Syllabus](#) [Readings](#) [Grading](#) [Projects](#) [Policies](#)

Social bias in text data

- Narratives [Field et al'19](#), [Field & Tsvetkov'19](#), [Park et al'20](#)
- Conversational domain [Breitfeller et al'19](#), [Field et al'20](#)

Social bias in NLP models & debiasing

- Embeddings [Manzini et al'19](#), [Kurita et al'19](#)
- Text classification [Jurgens et al'17a](#), [Xia et al'20](#)
[Kumar et al'19](#)

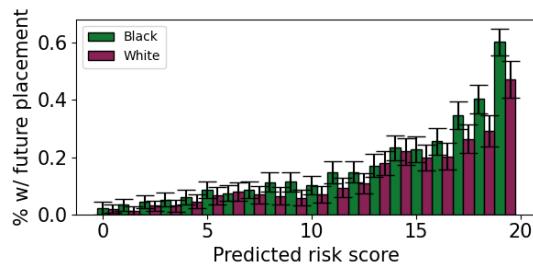
Fake news, misinformation

- Manipulation in narratives [Field et al'18](#)
- Factuality of automatically generated texts
[Pagnoni et al \(ongoing\)](#)

Privacy and profiling

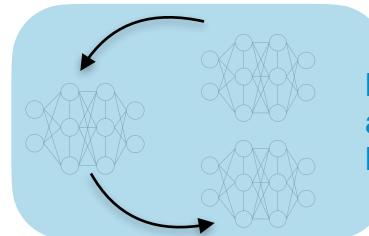
- [Jurgens et al'17b](#)

Partnerships with industry, government, and non-profit agencies

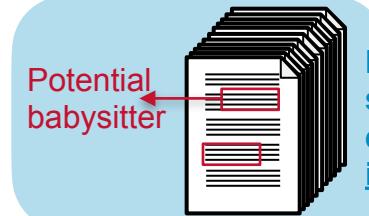


Opportunities and pitfalls of using NLP for predictive risk: a case study in the child welfare system

Ongoing Work [In submission to FAccT 2022]



Investigation of racial, gender, age, socioeconomic status biases in contact notes



Develop and analyze NLP systems for information extraction that prioritize interpretability and fairness

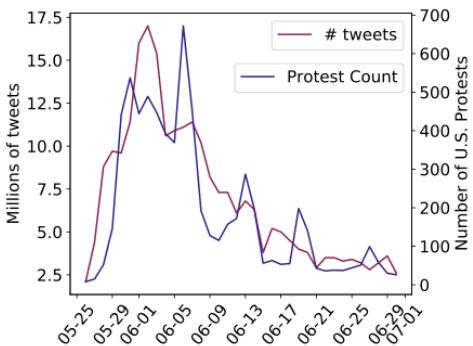


Continual investigation and critique of using automated tools in this setting

Future Work

Partnerships with industry, government, and non-profit agencies

Data for Black Lives

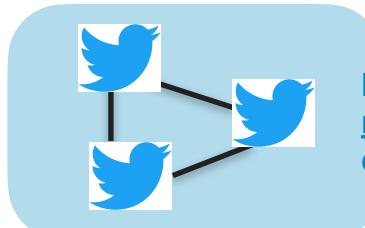


An Analysis of Emotions and the Prominence of Positivity in #BlackLivesMatter Tweets

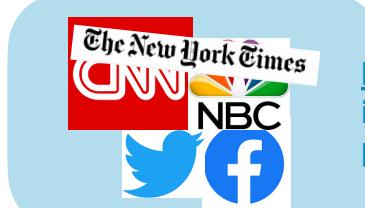
Ongoing Work [In revision for PNAS]



Fact Checker + Analysis
Video evidence of anti-black discrimination in China over coronavirus fears



Develop methods that integrate network analyses and NLP to characterize information spread



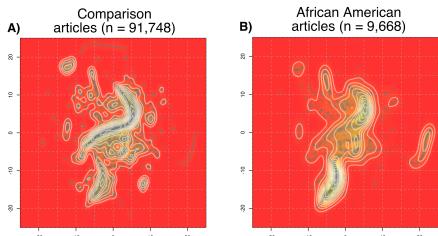
Multi-modal approaches for investigating how racism is perpetuated on public platforms



Continual investigation and critique of using automated tools in this setting

Future Work

Partnerships with industry, government, and non-profit agencies



“Controlled Analyses
of Social Biases in
Wikipedia Bios”
WebConf ‘22

English Wikipedia:
He *accepted* the option of injections of what was then
called stilboestrol.

Spanish Wikipedia:
Finalmente escogió las inyecciones de estrógenos.
Finally he *chose* estrogen injections.

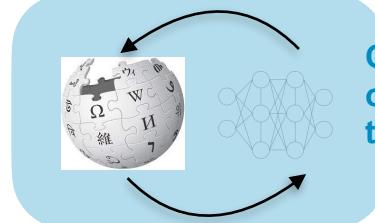
Russian Wikipedia:
Учёный предпочёл инъекции стильбэстрола
The scientist *preferred* stilbestrol injections.

“Multilingual
Contextual Affective
Analysis of LGBT
People Portrayals in
Wikimedia”
ICWSM ‘21

Ongoing Work



Automated methods for
identifying content gaps and
social biases



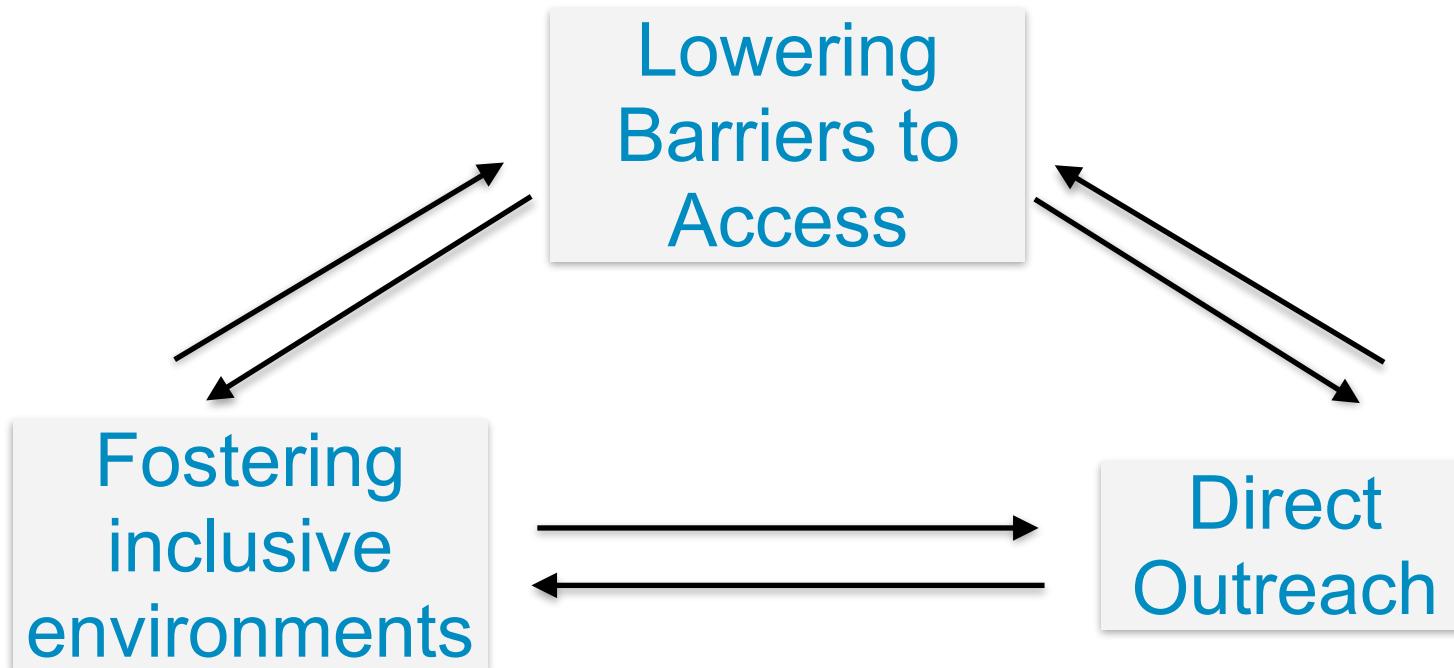
Controllable text generation and
output constraints in models
trained on Wikipedia data



Continual investigation and
critique of using automated
tools in this setting

Future Work

Improving Diversity in CS Research



Acknowledgements



Google Research



Natural Language
Processing



Political Africana
Science Studies

Network Science

Public Policy

Statistics

Economics

End