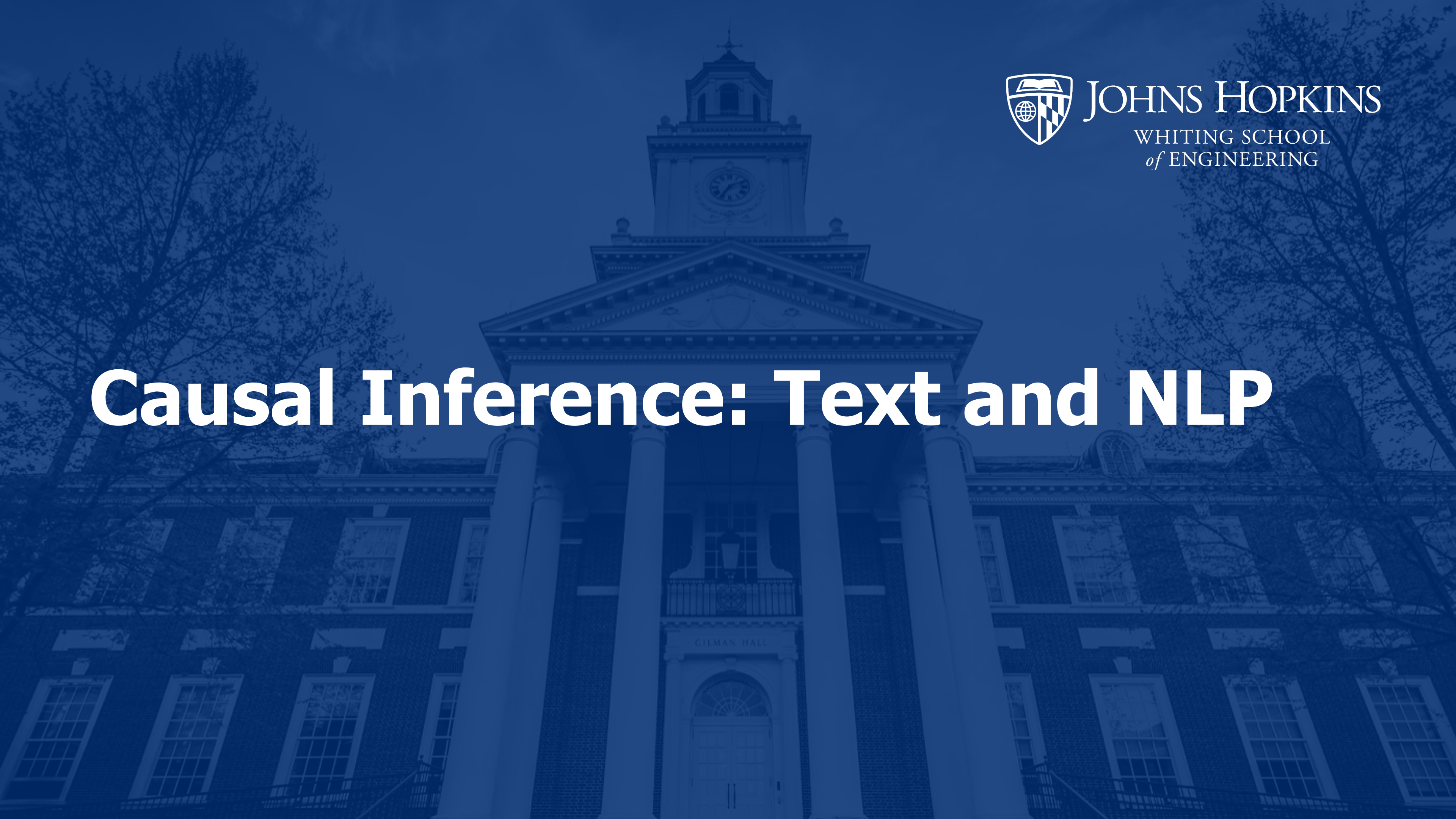
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Causal Inference: Text and NLP



Logistics

- HW 3 has been released and is due on Friday
  - On causal inference

- Midterm Exam
  - In class next Wednesday
  - Includes all material through Wednesday 3/6 (including guest/TA lectures and homeworks)
  - Sample problems released on Piazza
  - Review session Monday 3/11

Recap

- Methods for adjusting for confounders
  - Regression
  - Matching
  - Propensity scores (matching, weighting, and stratification)
- Confounders vs. Mediators vs. Colliders

Today: Causal Inference with Text

- Overview
- Adjusting for text as confounders (or mediators)
- Drawing from causal inference to improve NLP models

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Overview

What characteristics distinguish text from other data types?

- Text is high dimensional
  - Overfitting, violations of positivity

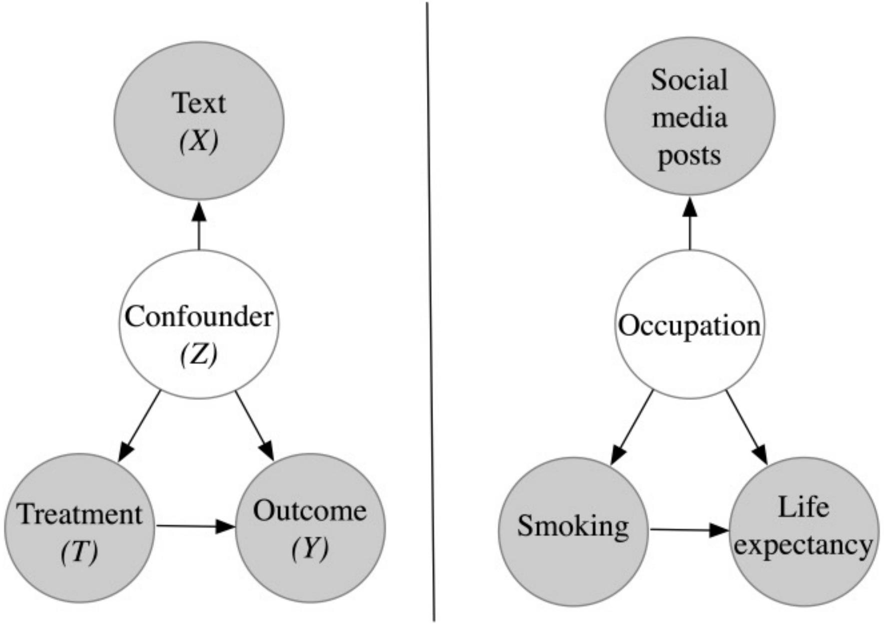
- Compared to other high dimensional data:
  - Text can be read and evaluated by humans
  - Designing meaningful representations of text is an open problem

Keith, Katherine, David Jensen, and Brendan O’Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." ACL. 2020.

Text as confounders

- Text data could either:
- (a) serve as a surrogate for potential confounders
- (b) the language of text itself could be a confounder

Example: the linguistic content of social media posts (confounder) could influence censorship (treatment) and future posting rates (outcome)



Keith, Katherine, David Jensen, and Brendan O’Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." ACL. 2020.

Text as treatment or outcome

- Do Wikipedia articles contain gender bias?
  - Treatment: Perceived gender
  - Outcome: Article text
  - Confounders/Mediators: Perceived characteristics other than gender

- Do Donald Trump's social media posts cause him to gain followers?
  - Treatment: Donald Trump's social media posts
  - Outcome: Donald Trump's follower counts
  - Confounders/Mediators: Changes in social media usage, current events

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Adjusting for text as confounders

Two similar approaches

- Topic Inverse Regression Matching
  - Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.
- “Causally sufficient” embeddings
  - Veitch, Victor, Dhanya Sridhar, and David Blei. "Adapting text embeddings for causal inference." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.

Adjusting for text as confounders: Topic Inverse Regression Matching

- Key ideas:
  - Matching (remember: direct or propensity) is a good approach for adjusting for text as confounder because analysts can manually evaluate the quality of the adjustment by comparing the matched treatment and control text
  - Most use cases what we need to match on are topics (as opposed to sentiment etc). We also may care about individual words
  - We need to match on aspects of the text that are predictive of treatment (definition of confounders)

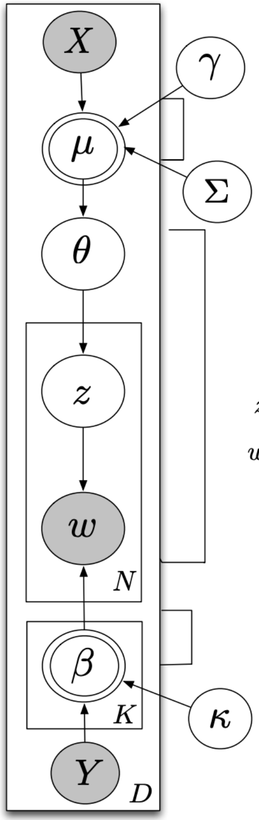
Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.

Topic Inverse Regression Matching using STM

- Define a function g(W) to create a low-dimensional estimate of confounding variables W

- Primary model for text representations: *structured topic model (STM)*

- LDA-style topic model that allows flexible inclusion of covariates



Topic Prevalence:

$$\mu_{d,k} = X_d \gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

Language Model:

$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

Topical Content:

$$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.

Step	Rationale
1. Estimate a structural topic model including the treatment vector as a content covariate.	Reduces dimension of the text
2. Extract each document's topics calculated as though treated (part of $g(\boldsymbol{W})$).	Ensures semantic similarity of matched texts
3. Extract each document's projection onto the treatment variable (part of $g(\boldsymbol{W})$).	Ensures similar treatment probability of matched texts
4. Use a low-dimensional matching method to match on $g(\boldsymbol{W})$ and estimate treatment effects using the matched sample.	Standardizes matching

Example application: Effects of Censorship on Chinese social media

- “Is censorship completely determined by the text of a particular post, or does censorship become more targeted toward users based on their previous censorship history?”

- Methods:
  - Use TIRM to identify pairs of nearly identical social media posts written by nearly identical users, where one is censored and the other is not
  - Examine subsequent censorship rates of each user
  - [Also examine rate of posting after censorship]

Example application: Effects of Censorship on Chinese social media

- "Is censorship completely determined by the text of a particular post, or does censorship become more targeted toward users based on their previous censorship history?"

- Results:
  - Having a post censored increases the probability of future censorship significantly
  - It does not decrease number of future posts by the censored user

- Conclusions:
  - Option 1: algorithmic targeting of censorship, where social media users are more likely to be censored after censorship because they are flagged by the censors
  - Option 2: social media users may chafe against censorship and respond by posting increasingly sensitive content that is more likely to be censored

A different method: develop “causally sufficient” text embeddings

- Text is high dimensional and data is finite: difficult to fit models directly to text
- Instead, “reduce the text to a low-dimensional representation that suffices for causal identification and enables efficient estimation from finite data.”

- Two key ideas:
  - Supervised dimensionality reduction: we don’t need the full text, causal inference only requires the parts of text that are predictive of the treatment and outcome
  - Efficient language modeling: design representations of text to dispose of “linguistically irrelevant information”, presumed to also be “causally irrelevant”

- [They also do a variant based on a topic model]

Veitch, Victor, Dhanya Sridhar, and David Blei. "Adapting text embeddings for causal inference." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.

General approach: develop “causally sufficient” text embeddings

- Start with a language model (BERT) and modify it to produce 3 outputs:
  - 1) document-level embeddings
  - 2) a map from the embeddings to treatment probability
  - 3) a map from the embeddings to expected outcomes for the treated and untreated
  - [(2) and (3) are small added neural networks on the original model]

- [They also do a variant based on a topic model]

General approach: develop "causally sufficient" text embeddings

- Train model to predict outcome, treatment, and with language-modeling objective (e.g. to learn meaningful text representations)

$$\begin{aligned} L(\mathbf{w}_i;\xi,\gamma) = &\left(y_i - \tilde{Q}(t_i,\lambda_i;\gamma)\right)^2 \longrightarrow \text{Outcome} \\ &+ \mathsf{CrossEnt}\left(t_i, \tilde{g}(\lambda_i;\gamma)\right) \longrightarrow \text{Treatment} \\ &+ L_{\mathrm{U}}(\mathbf{w}_i;\xi,\gamma). \longrightarrow \text{Language modeling} \end{aligned}$$

- Use propensity score estimates (g) to compute causal estimands

Evaluation

- Two settings:
  - Peer-reviewed journal articles: Causal effect of including a theorem on paper acceptance.
    - Treatment: the word “theorem” occurs in the paper
    - Confounder: article abstract (subject of the paper)
    - Outcome: accept/reject
  - Effect of gender on Reddit popularity
    - Treatment: “male” label
    - Mediator: Post text (topic or style)
    - Outcome: Popularity score

How can we use this data for *evaluation* rather than analysis?

Evaluations

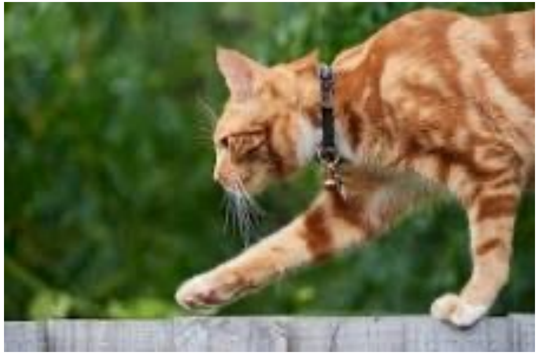
- Simulated data:
  - Use real confounders and treatments
  - Simulate outcomes (so we know the “true” causal effect)
- Their findings:
  - 1) Yes, language modeling helps recover simulated effects
  - 2) Yes, supervised dimensionality helps
  - 3) Their proposed models C-BERT and C-ATM outperform alternatives

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Drawing from Causal Inference to Improve NLP models

Drawing from Causal Inference to Improve NLP models

- ML in general typically captures associates, not causal effects
- Models are prone to overfitting, exploit spurious correlations in the data
  - E.g. train a model to identify photos of dogs from cats; Model learns that dogs always have collars



“DOG”

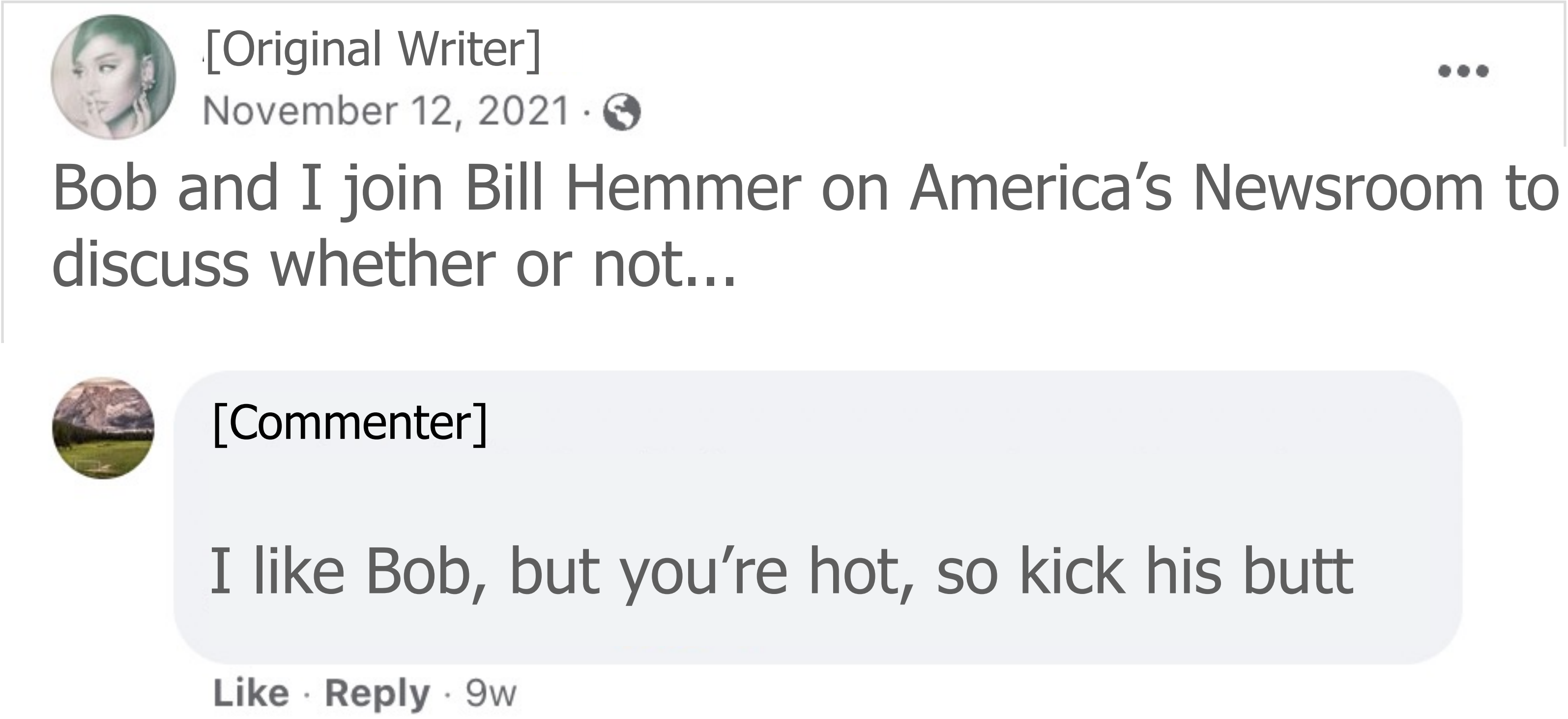
Drawing from Causal Inference to Improve NLP models

- ML in general typically captures associates, not causal effects
- Models are prone to overfitting, exploit spurious correlations in the data
  - E.g. train a model to identify photos of dogs from cats; Model learns that dogs always have collars

- Maybe by drawing from causal inference we can train models to ignore these spurious correlations, especially for tasks where it's hard to collect good training data
- Case study: drawing from causal inference to detect *subtle gender bias*

Field, Anjalie, and Yulia Tsvetkov. "Unsupervised Discovery of Implicit Gender Bias." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Goal: Identify text containing (subtle) gender bias



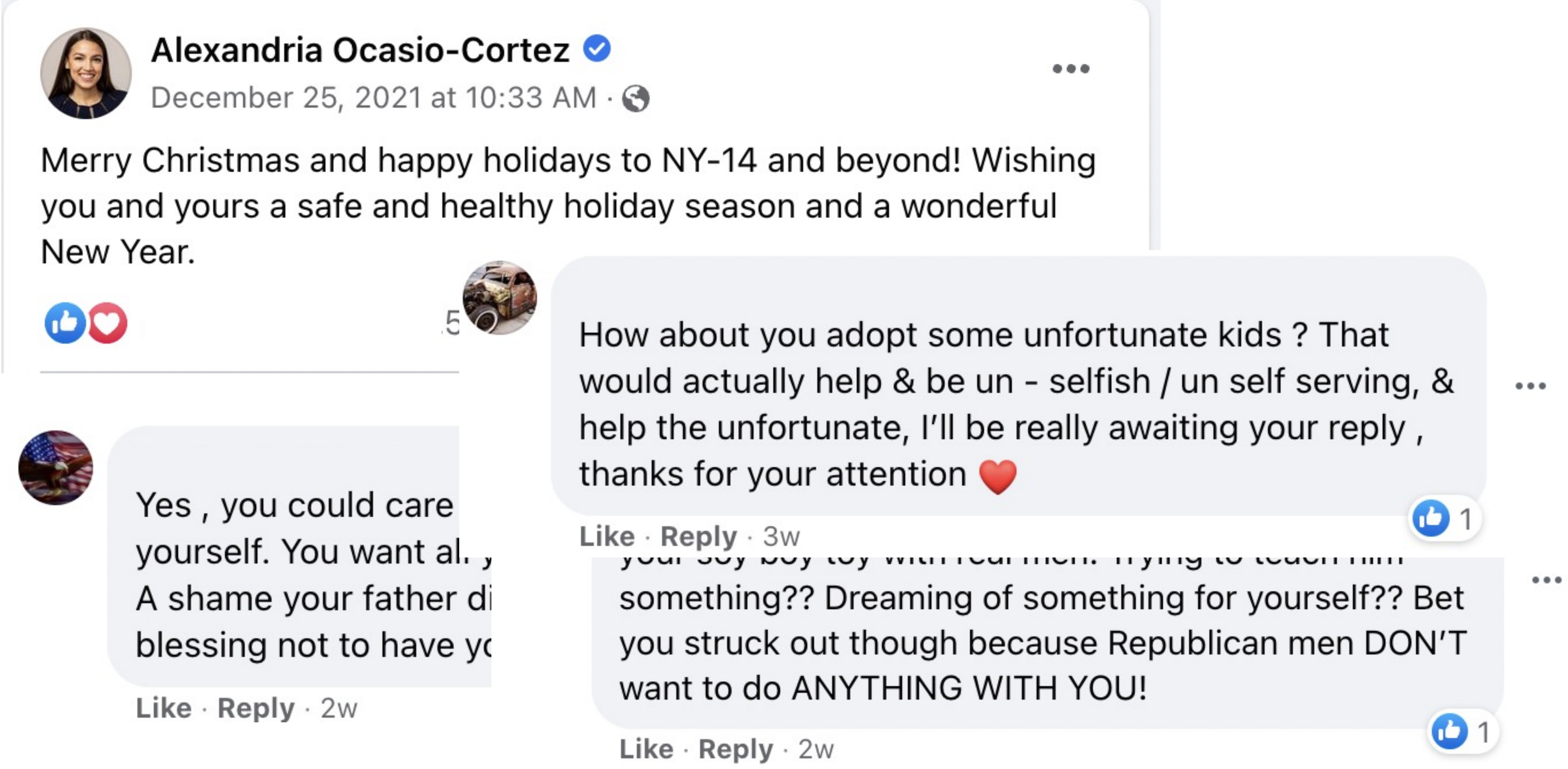
[Original Writer]

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



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.

5

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 ❤️

Like · Reply · 3w · 1

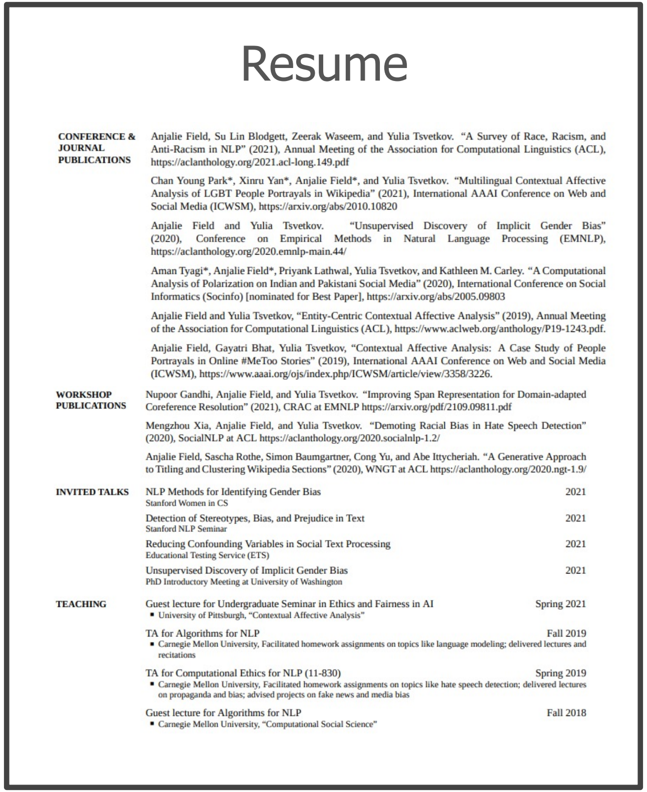
Yes , you could care
yourself. You want al. y
A shame your father di
blessing not to have yo

Like · Reply · 2w

your soy boy toy with real men. Trying to teach him 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 · 1

(Cuthbertson et al., 2019; Di Meco and Brechenmacher, 2020;2019)



Resume

CONFERENCE & JOURNAL PUBLICATIONS

Anjalie 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

Chan Young Park*, Xinru Yan*, Anjalie 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

Anjalie 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*, Anjalie 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

Anjalie Field and Yulia Tsvetkov, “Entity-Centric Contextual Affective Analysis” (2019), Annual Meeting of the Association for Computational Linguistics (ACL), https://www.aclweb.org/anthology/P19-1243.pdf.

Anjalie Field, Gayatri Bhat, 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/3358/3226.

WORKSHOP PUBLICATIONS

Nupoor Gandhi, Anjalie Field, and Yulia Tsvetkov. “Improving Span Representation for Domain-adapted Coreference Resolution” (2021), CRAC at EMNLP https://arxiv.org/pdf/2109.09811.pdf

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. “Demoting Racial Bias in Hate Speech Detection” (2020), SocialNLP at ACL https://aclanthology.org/2020.socialnlp-1.2/

Anjalie Field, Sascha Rothe, Simon Baumgartner, Cong Yu, and Abe Ittycheriah. “A Generative Approach to Titling and Clustering Wikipedia Sections” (2020), WNGT at ACL https://aclanthology.org/2020.ngt-1.9/

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”

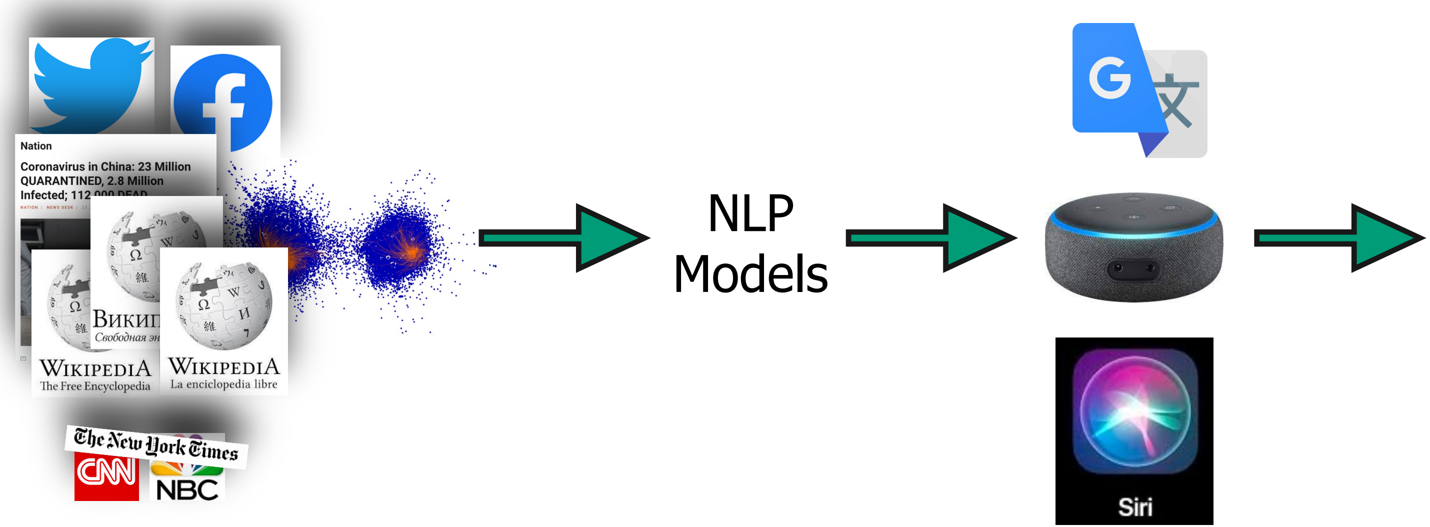
TA for Algorithms for NLP Fall 2019
- Carnegie Mellon University, Facilitated homework assignments on topics like language modeling; delivered lectures and recitations

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?



NLP Models

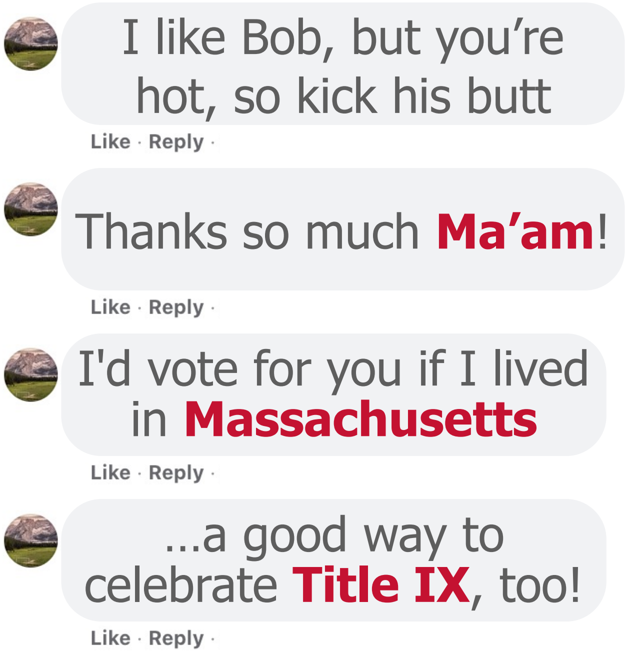
"Oh, you work at an office? I bet you're a secretary"

"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**!

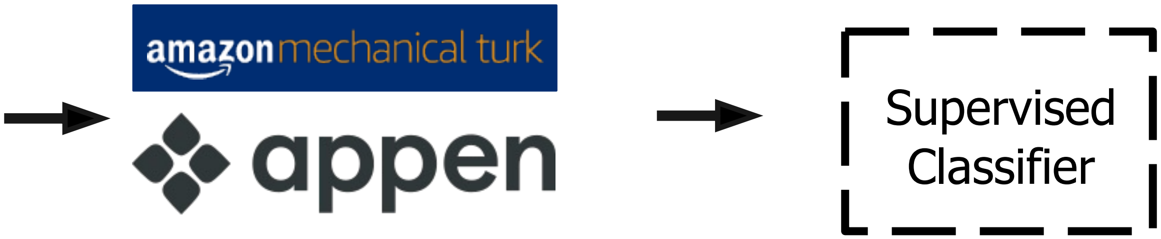
Like · Reply ·

I'd vote for you if I lived in **Massachusetts**

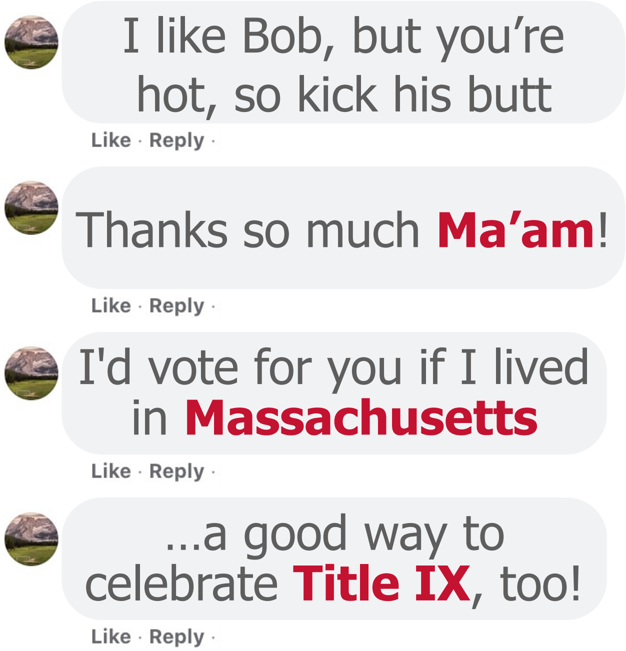
Like · Reply ·

…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**!

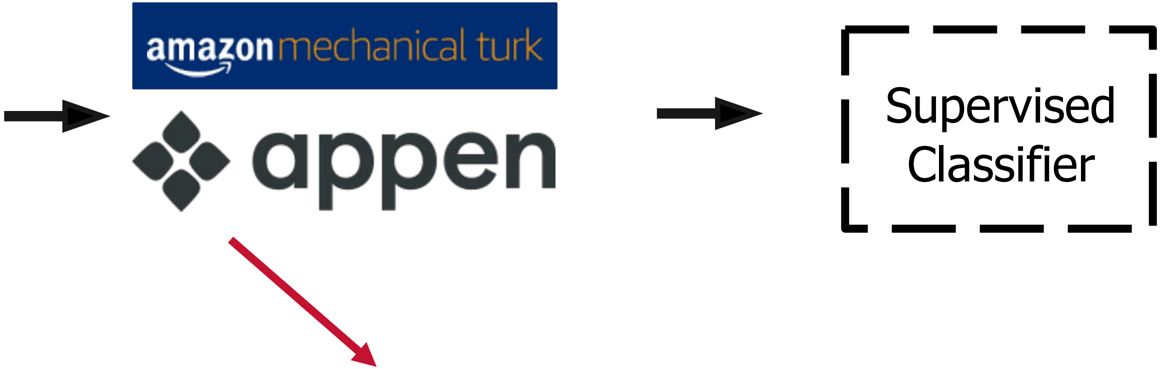
Like · Reply ·

I'd vote for you if I lived in **Massachusetts**

Like · Reply ·

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

Like · Reply ·

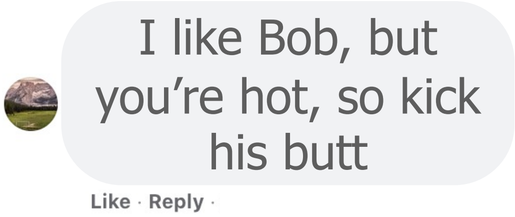


Supervised Classifier

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

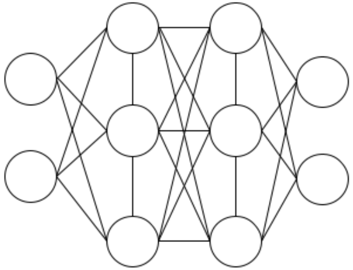
Proposed approach: Comments contain gender bias if they are highly predictive of gender

- 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**!

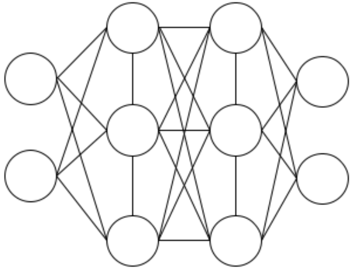
Like · Reply ·

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

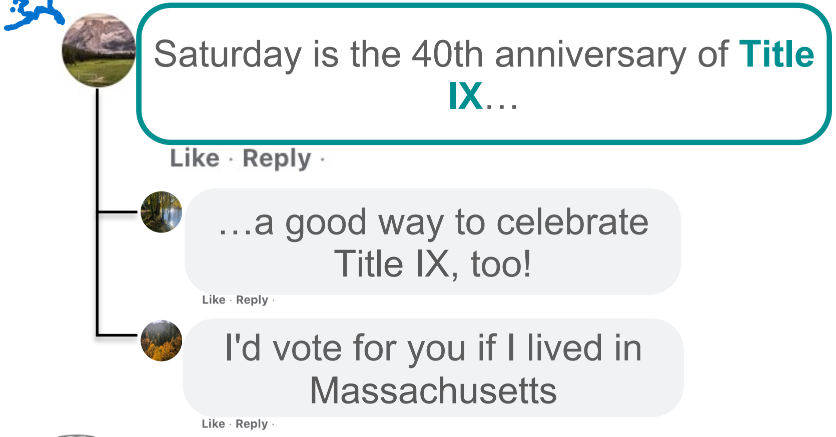
→ Addressed to Woman

→ Addressed to Woman

→ Addressed to Woman

Challenge: Text main contain *confounds* 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**…

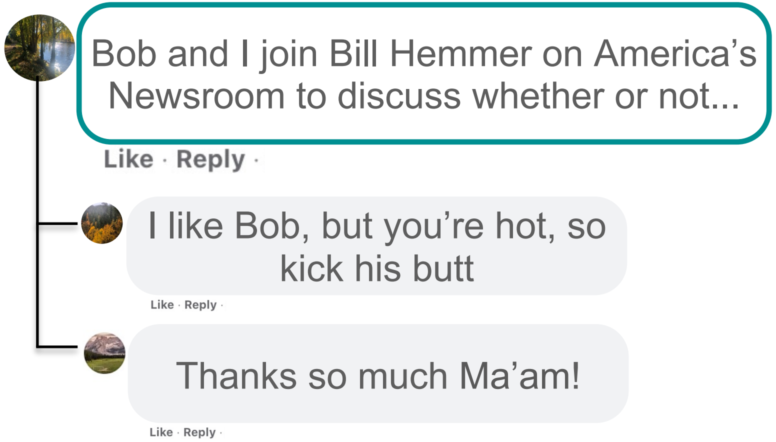
Like · Reply ·

…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 ·

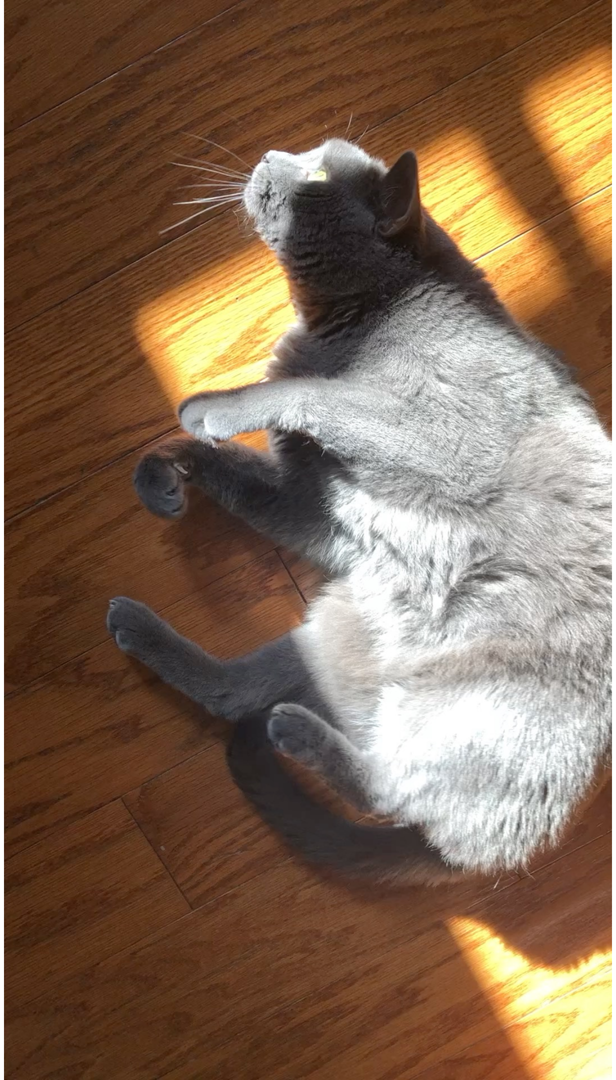
I like Bob, but you're hot, so kick his butt

Like · Reply ·

Thanks so much Ma'am!

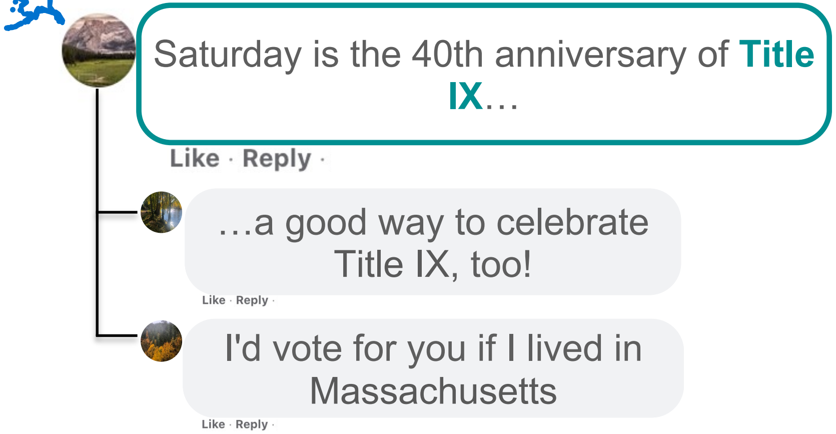
Like · Reply ·

Break



Challenge: Text main contain *confounds* 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**…

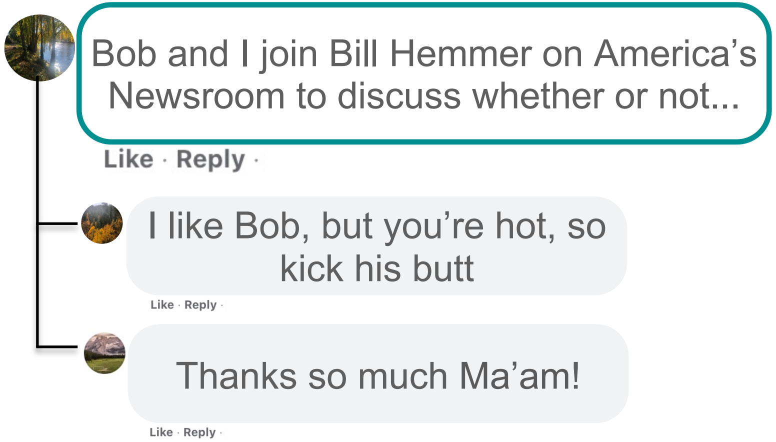
Like · Reply ·

…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 ·

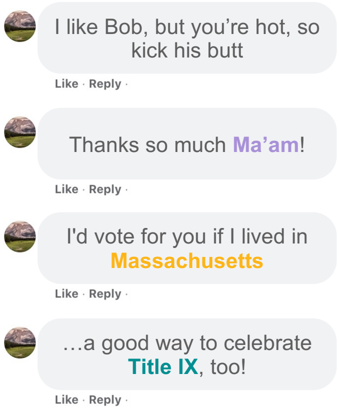
Thanks so much Ma'am!

Like · Reply ·

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

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



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

Like · Reply ·

Thanks so much **Ma'am**!

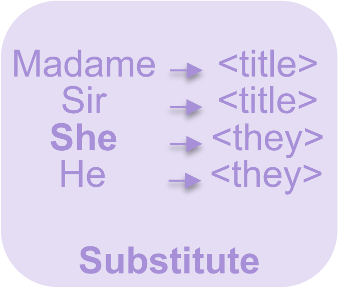
Like · Reply ·

I'd vote for you if I lived in **Massachusetts**

Like · Reply ·

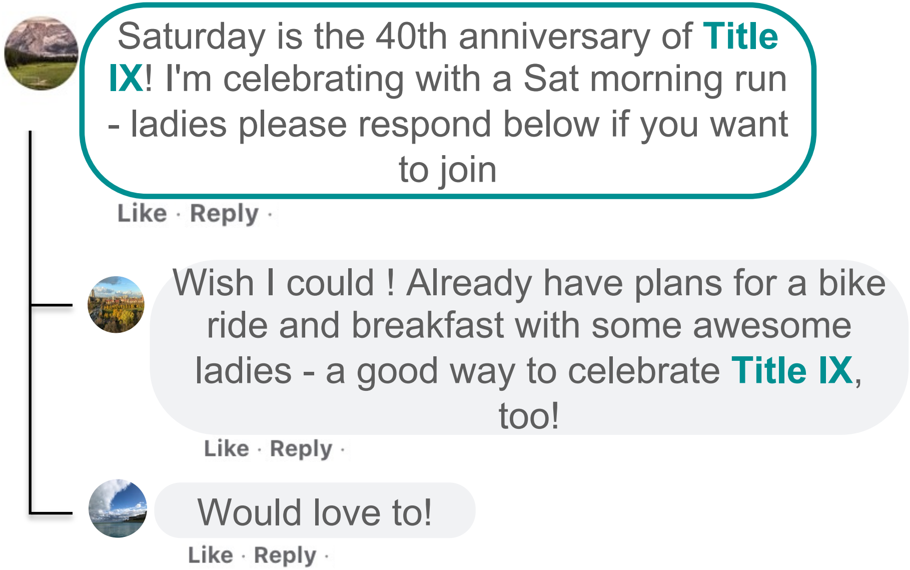
…a good way to celebrate **Title IX**, too!

Like · Reply ·



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 ·

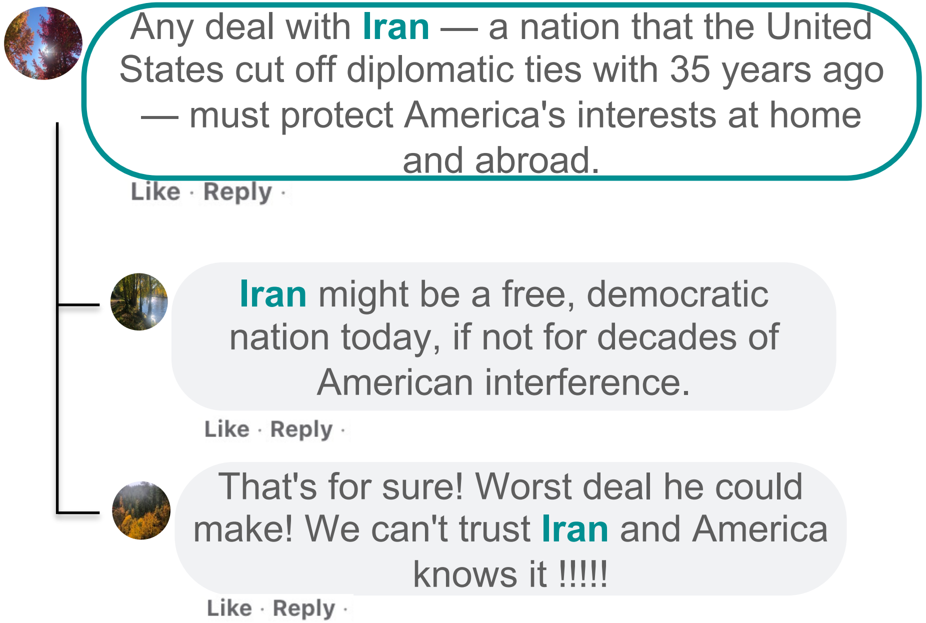
Wish I could ! Already have plans for a bike ride and breakfast with some awesome ladies - a good way to celebrate **Title IX**, too!

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 and abroad.

Like · Reply ·

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

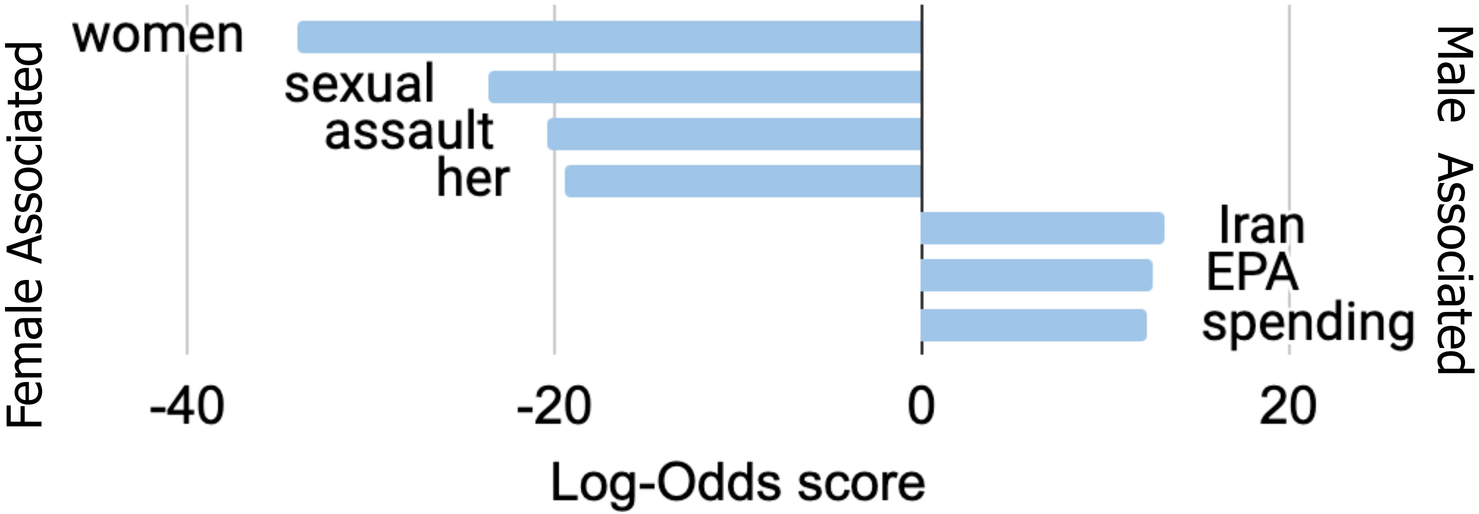
Like · Reply ·

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

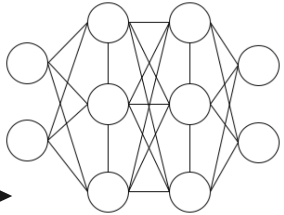
~~Saturday is the 40th anniversary of **Title IX**! I'm celebrating with a Sat morning run - ladies please respond below if you want to join~~

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 developing a nuclear weapon.



Text classifier to predict WRITER_GENDER

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

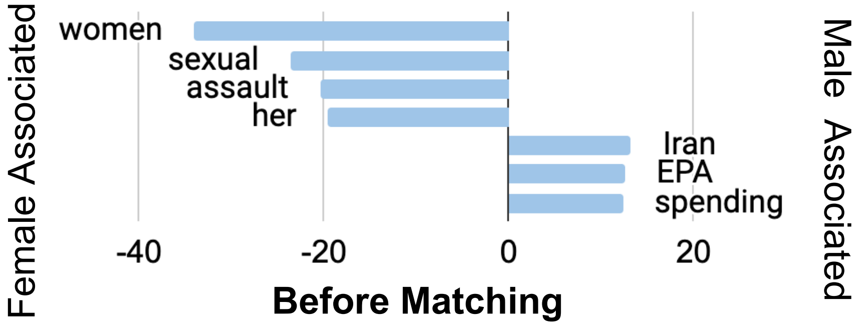
$$e_i = P(W.Gender_i = F|Post_i) \approx 0.91$$

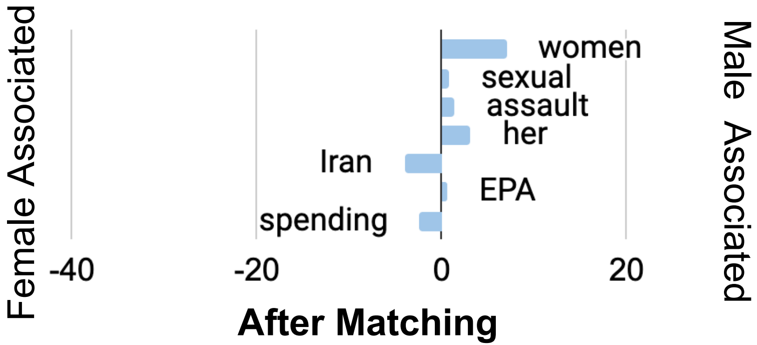
$$e_j = P(W.Gender_j = F|Post_j) \approx 0.33$$

$$e_k = P(W.Gender_k = F|Post_k) \approx 0.32$$

$$argmin_j|e_k - e_j|$$

Propensity matching for *observed* confounding variables

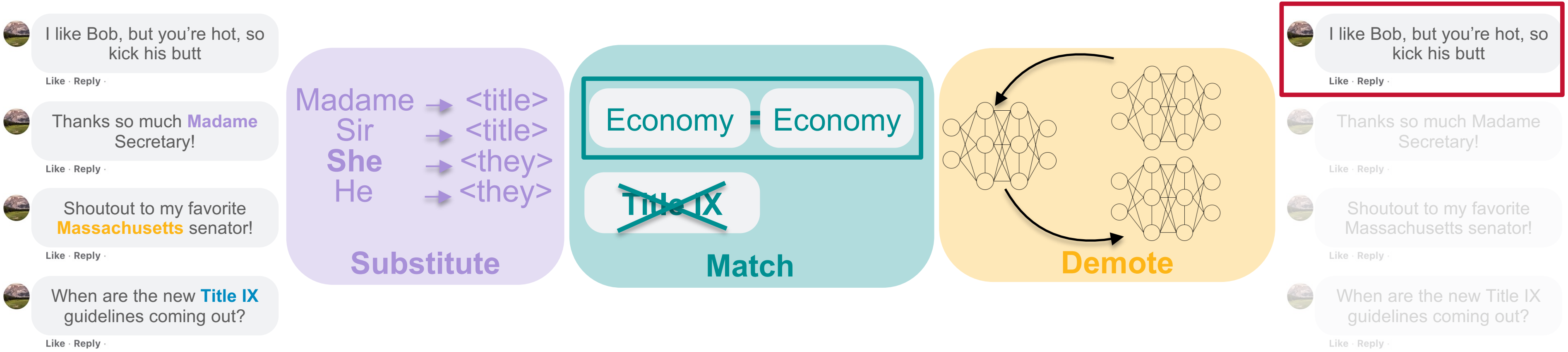




Propensity matching breaks associations between gender and context in the training data

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



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

Like · Reply ·

Thanks so much Madame Secretary!

Like · Reply ·

Shoutout to my favorite Massachusetts senator!

Like · Reply ·

When are the new Title IX guidelines coming out?

Like · Reply ·

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

Like · Reply ·

Thanks so much Madame Secretary!

Like · Reply ·

Shoutout to my favorite Massachusetts senator!

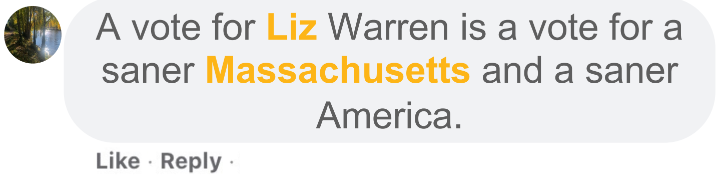
Like · Reply ·

When are the new Title IX guidelines coming out?

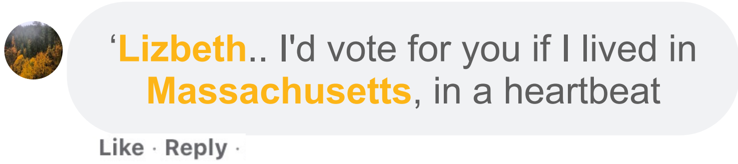
Like · Reply ·

Adversarial training for *latent* confounding variable

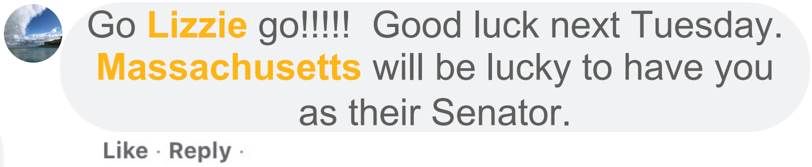
- Comments may references traits of the addressee (such as occupation, nationality, nicknames, etc.) that are correlated with gender
- Difficult to enumerate all of them
- Often unique to individuals (difficult to make matches)



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

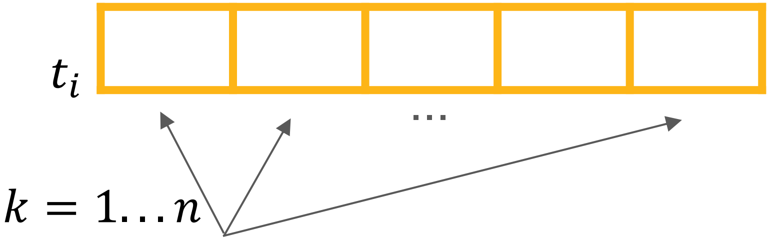


'**Lizbeth**.. I'd vote for you if I lived in **Massachusetts**, in a heartbeat



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

Represent latent confounding variables as a vector



$$p(addressee = k|comment) \propto p(addressee = k)p(comment|addressee)$$
$$= p(addressee = k) \prod_{w_i \in comment} p(w_i|k)$$

[word-level probability estimates are derived from log-odds with a Dirichlet prior scores]

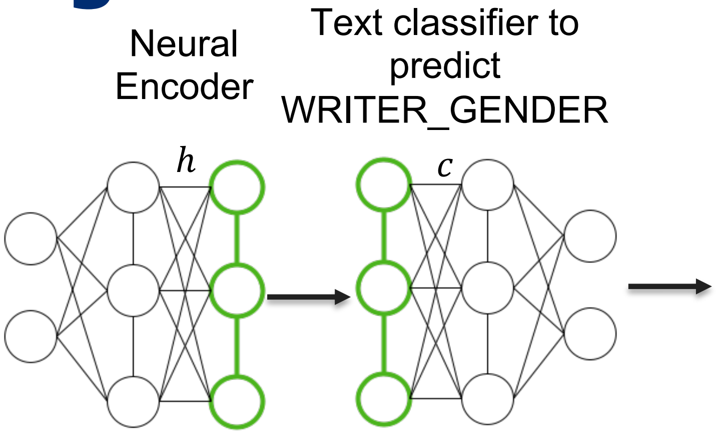
Adversarial training for *latent* confounding variables

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

Like · Reply ·

Shoutout to my favorite **Massachusetts** senator!

Like · Reply ·



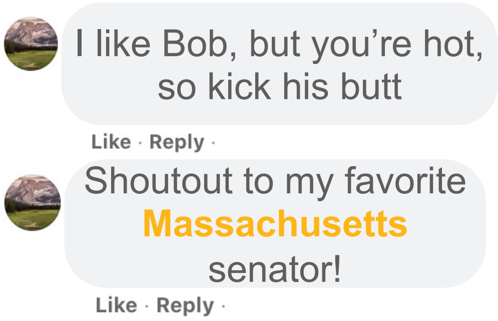
$P(W.Gender = F) = 0.90$

$CE(c(h(x_i)), y_i)$

$P(W.Gender = F) = 0.90$

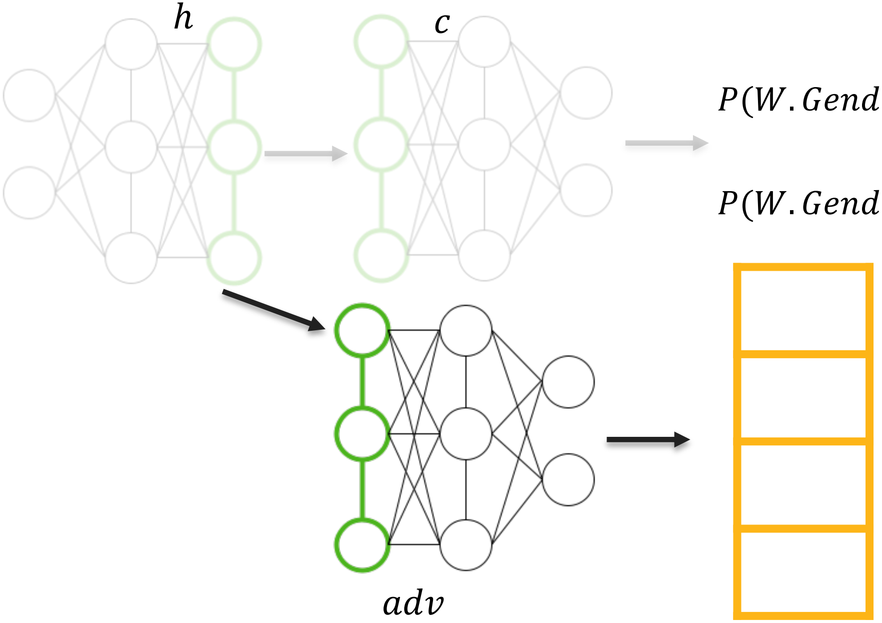
(Kumar et al. 2019)

Adversarial training for *latent* confounding variables



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

Shoutout to my favorite **Massachusetts** senator!



Neural Encoder

Text classifier to predict WRITER_GENDER

h

c

$P(W.Gender = F) = 0.90$

$P(W.Gender = F) = 0.90$

$CE(adv(h(x_i)), t_i)$

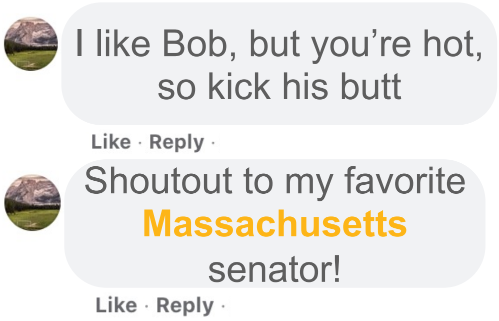
adv

Adversary

t_i

Vector representation of latent traits

Adversarial training for *latent* confounding variables

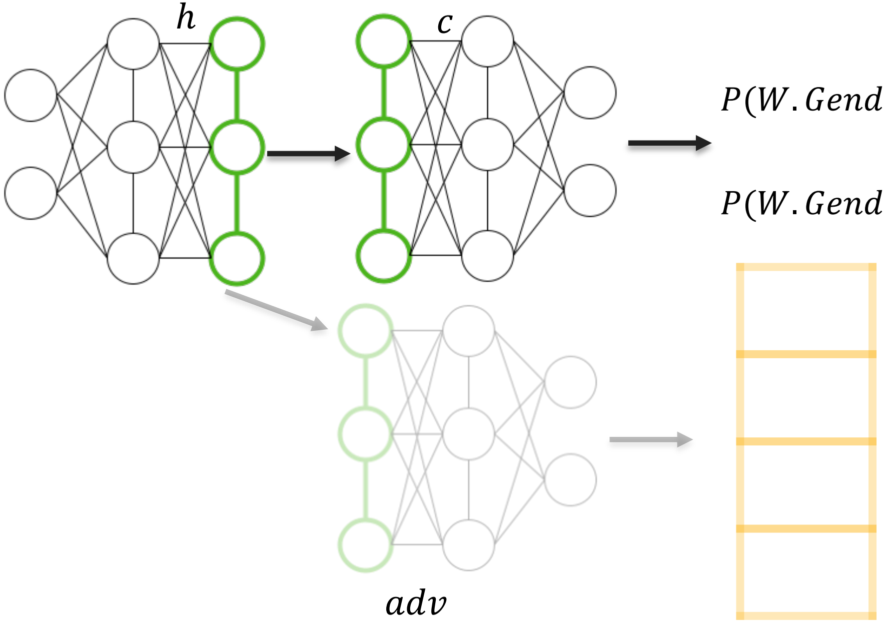


I like Bob, but you’re hot, so kick his butt

Like · Reply ·

Shoutout to my favorite **Massachusetts** senator!

Like · Reply ·



Neural Encoder

Text classifier to predict WRITER_GENDER

h

c

$P(W.Gender = F) = 0.90$

$CE(c(h(x_i)), y_i)$

$P(W.Gender = F) = 0.90$

$CE(adv(h(x_i)), U)$

adv

Adversary

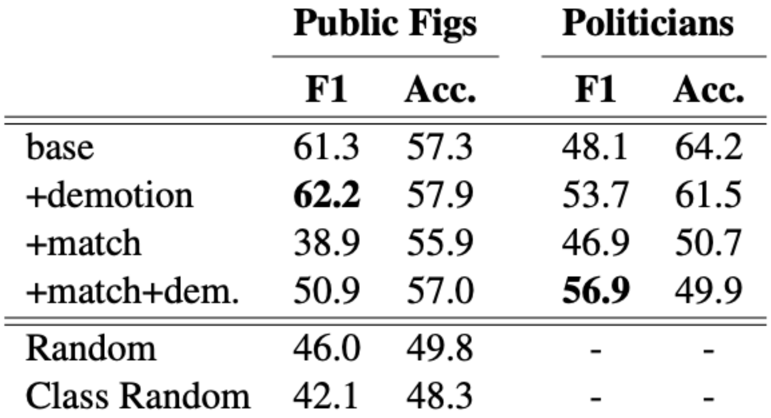
t_i

Vector representation of latent traits

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

Self-reported microaggressions

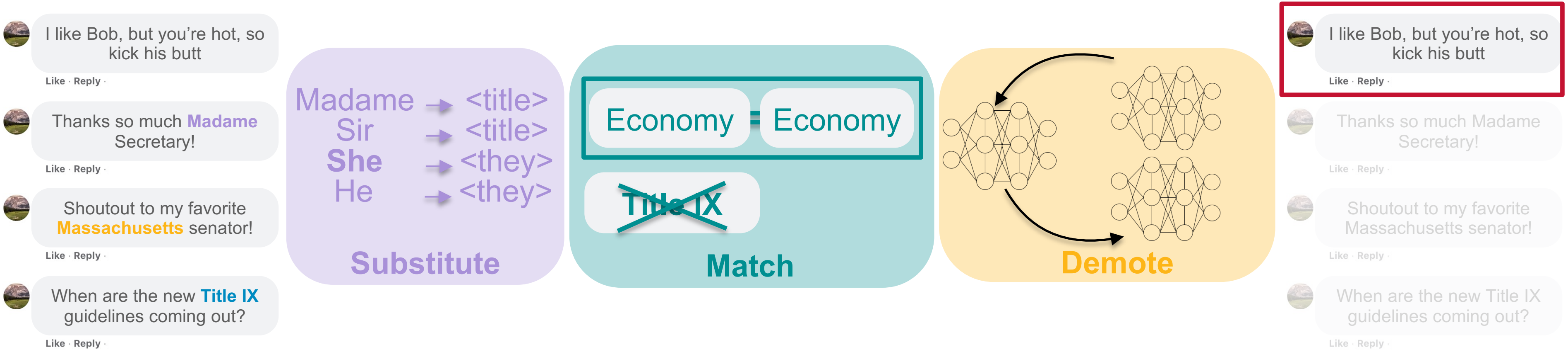


	Public Figs		Politicians	
	F1	Acc.	F1	Acc.
base	61.3	57.3	48.1	64.2
+demotion	**62.2**	57.9	53.7	61.5
+match	38.9	55.9	46.9	50.7
+match+dem.	50.9	57.0	**56.9**	49.9
Random	46.0	49.8	-	-
Class Random	42.1	48.3	-	-

- Models are not trained at all for this task; they are only trained for gender-of-addressee prediction, but they still perform better than chance

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



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

Like · Reply ·

Thanks so much Madame Secretary!

Like · Reply ·

Shoutout to my favorite Massachusetts senator!

Like · Reply ·

When are the new Title IX guidelines coming out?

Like · Reply ·

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

Like · Reply ·

Thanks so much Madame Secretary!

Like · Reply ·

Shoutout to my favorite Massachusetts senator!

Like · Reply ·

When are the new Title IX guidelines coming out?

Like · Reply ·

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.”

Recap

- Overview:
  - Text as confounders, treatment, or outcome

- Text as confounders
  - Topic modeling and language modeling to adjust for text

- Drawing from causal inference to improve NLP models
  - Applying ideas from causal inference to model development

- Next class:
  - Network Analysis

References

- Keith, Katherine, David Jensen, and Brendan O’Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." ACL. 2020.
- Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.
- Veitch, Victor, Dhanya Sridhar, and David Blei. "Adapting text embeddings for causal inference." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.
- Field, Anjalie, and Yulia Tsvetkov. "Unsupervised Discovery of Implicit Gender Bias." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.