arxiv.org/abs/1905.12741

Adapting Text Embeddings for Causal Inference

Victor Veitch*, **Dhanya Sridhar***, David Blei
Columbia University

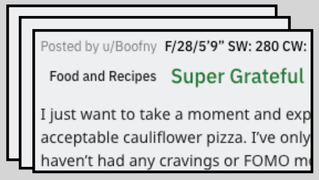
Motivating example: What is the effect of perceived gender on popularity of post?

T: perceived gender is male Y: number of upvotes

$T = 1$

$\mathbb{E}[Y \mid T = 1] = 50$

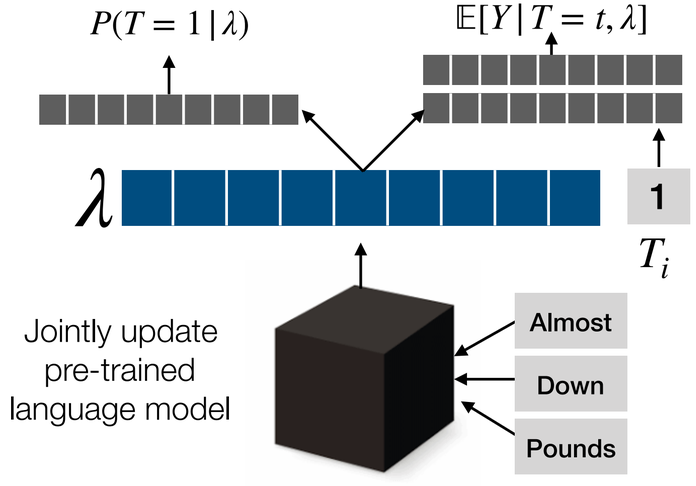
$T = 0$



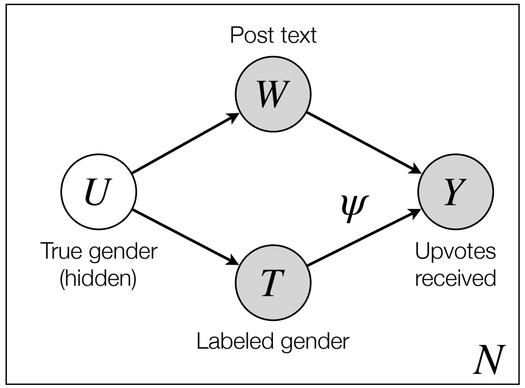
$\mathbb{E}[Y \mid T = 0] = 20$

Difference in upvotes between groups explained by different writing styles between women and men

Causal BERT — an example



Use predicted values in plug-in estimators



$$\psi = \mathbb{E}[Y \mid T = 1; \mathrm{do}(\mathrm{T} = 1)] - \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = 1; \mathrm{do}(\mathrm{T} = 0)]$$

Average effect of treatment on the treated (ATT)
Counterfactual outcome for post written by people self-labeled as men

Problem: no ground truth ATTs, and no way to realistically simulate text

Strategy:

- Choose observed covariate that can be predicted from text. E.g., subreddit
- Simulate outcomes that depend on effect from treatment and effect from covariate (confounding)
- E.g., linear model:

$$Y_i = T_i + \beta_1(\pi(Z_i) - 0.5) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0,\gamma)$$

Proportion of self-labeled men in subreddit Z

Problem: Can write ATT in terms of observed variables but, need to adjust for text of the document

1. **Insight 1:** for causal adjustment, the part of text which carries information about treatment and outcome is all that matters.
2. **Insight 2:** outcomes are produced by humans processing text. Thus, only text information that carries natural language meaning matters.
3. **Taken together:** supervised dimensionality reduction + language modeling

Top-level comments from subreddits with gender labels and upvotes

Noise:	$\gamma = 1.0$			$\gamma = 4.0$		
Confounding:	Low	Med.	High	Low	Med.	High
Ground truth	1.00	1.00	1.00	1.00	1.00	1.00
Unadjusted	1.03	1.24	3.48	0.99	1.22	3.51
NN $\hat{\psi}^{Q}$	1.03	1.18	2.04	0.89	1.08	2.24
NN $\hat{\psi}^{\mathrm{plugin}}$	1.03	1.18	1.40	0.85	1.05	2.07
C-ATM $\hat{\psi}^{Q}$	1.01	1.16	2.45	1.04	1.04	1.72
C-ATM $\hat{\psi}^{\mathrm{plugin}}$	1.01	1.13	2.09	0.95	0.94	1.11
C-BERT $\hat{\psi}^{Q}$	1.07	1.07	1.14	1.50	0.95	1.12
C-BERT $\hat{\psi}^{\mathrm{plugin}}$	1.08	1.15	0.94	2.07	1.07	1.27

NN: Feedforward neural network with bag-of-words (NN)
C-BERT: Causal BERT [this paper]
C-ATM: Causal supervised topic model with amortized inference [this paper]