Text Analysis for Social Sciences in Python

Week 13: Causal Inference with Text Data

WINTER SEMESTER 21/22

HUYEN NGUYEN

HTTP://HUYENTTNGUYEN.COM

Last week…

Overview introduction/refresher of causal inference framework

If you haven't yet, kindly take 10' to fill in the feedback survey – your inputs are hugely appreciated!
https://evasys-online.uni-hamburg.de/evasys/online.php?pswd=1S46L

Today's agenda

- Text as Confounders
- Text as Outcomes
- Text as Treatment
- Text as Mediator
- Exercise: Causal Interpretation and Correlation

Example 1

An online forum has allowed its users to indicate their preferred gender in their profiles with an icon.

They notice that users who label themselves with the "woman" icon tend to receive fewer “likes“ Y on their posts W.

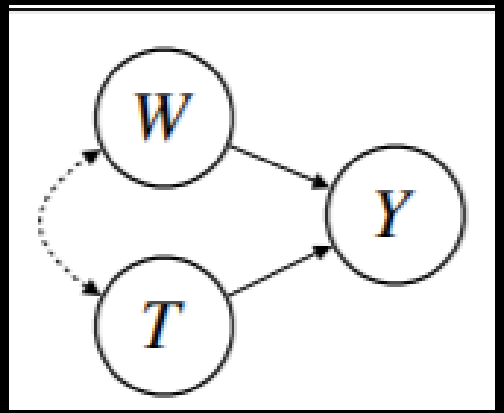
To better evaluate their policy of allowing gender information in profiles, they ask: Does being perceived as a woman cause a decrease in popularity for a post?

Example 1(cont)

What is the treatment?

What is the outcome?

What is the counterfactual question?



Example 1(cont)

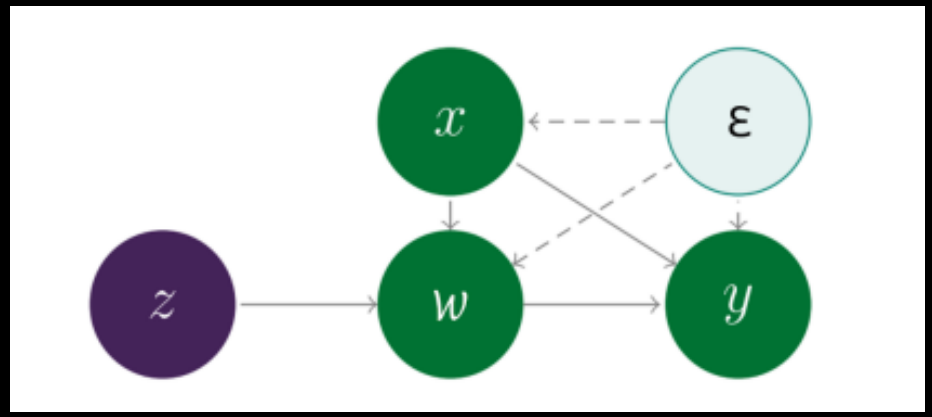
What is the treatment? Causal effect of being perceived as a woman T.

What is the outcome? The likes a post receive Y.

What is the counterfactual question?

If we could manipulate the gender icon on a given post, how many likes would the post have received?

The Empirical Problem



y = outcome
w = (text) treatments
x = observable covariates
ε = confounders
z = instruments

(!) spurious correlation by confounders i.e. variables that are correlated with both the treatment and outcome.

What are the likely (text) confounders ε in Example 1?

Example 2

A medical research center wants to build a classifier $\widehat{Y}(W)$ to detect clinical diagnoses Y from the textual narratives W of patient medical records. The records are aggregated across multiple hospital sites, which vary both in the frequency of the target clinical condition and the writing style ε of the narratives.

When the classifier is applied to records from sites that were not in the training set, its accuracy decreases. Post-hoc analysis indicates that it puts significant weight on seemingly irrelevant features, such as formatting markers.

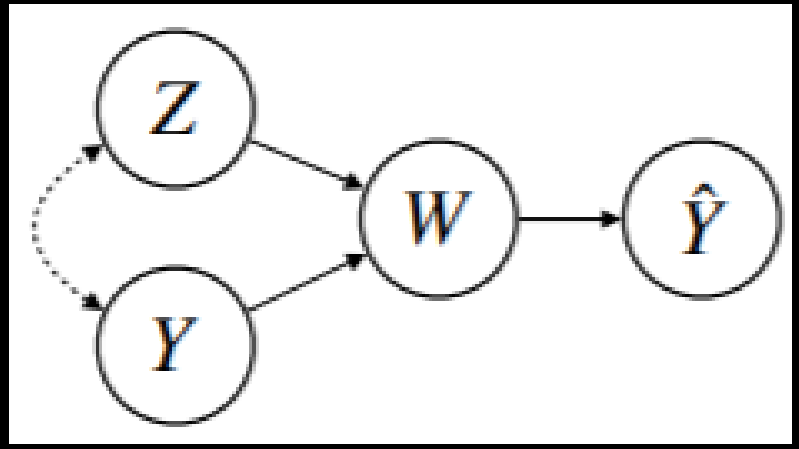
→ The counterfactual question: Does the classifier's prediction change if we intervene to change the hospital site, while holding the true clinical status fixed?

Example 2 (cont)

Goal: get the classifier to rely on phrases that express clinical facts, NOT writing style.

BUT, in the training data, the clinical condition and the writing style are spuriously correlated.

What are the possible confounding variables?



Pitfalls of NLP

NLP systems are often black boxes → hard to understand how human interpretable features of the text lead to the observed predictions.

In this setting, we want to know if some part of the text (e.g., some sequence of tokens) causes the output of an NLP method (e.g., classification prediction).

Assumptions of Causal Inference

Ignorability: $\big(Y(1), Y(0)\big) \perp T$

Positivity: $0 < \Pr(T = 1 | X = x) < 1$

Consistency: $\forall i \; s.t \; T_i = t, Y_i(t) = Y_i$

CI with text as confounders

Where? Observational data sets with texts

Challenge: How to condition on the text such that it blocks confounding with NLP methods.

CI with text as confounders (cont)

Where? Observational data sets with texts

Challenge: How to condition on the text such that it blocks confounding with NLP methods.

→ Unsupervised dim reduction techniques. E.g. topic models (Roberts et al. 2020), embedding methods, auto-encoders

CI with text as confounders (cont)

Where? Observational data sets with texts

Challenge: How to condition on the text such that it blocks confounding with NLP methods.

→ Unsupervised dim reduction techniques. E.g. topic models (Roberts et al. 2020), embedding methods, auto-encoders

→ Supervised models by learning low-dimensional variables that can predict treatment and outcome well, confounding properties can be found in text.

. E.g: Veitch, Sridhar and Blei (2020) adapted pre-trained language models and supervised topic models to predict the treatment and outcome.

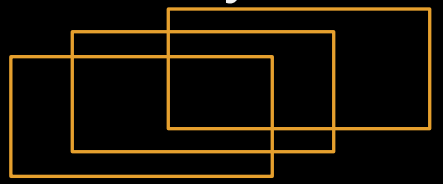
CI with text as outcomes

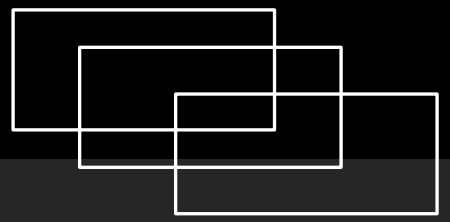
Where? Open-ended responses from randomized experiments

Challenge:

Distill high-dimensional data into low-dimensional quality of interest

- E.g: The effect of intensive English skill course intervention on readability in student essays





Readability (1)

Readability (0)

Readability is a latent variable of language!
→ Unknown outcome measure at the start of the experiment

CI with text as treatments

Where?

1) Study of treatment discovery:

e.g: topics &latent dimensions of candidate biographies that drove voter evaluations Fong & Grimmer 2016

n-gram influential writing style in marketing materials that increase sales figures Pryzant et al., 2017)

→producing interpretable features of the text that can be causally linked to outcomes.

CI with text as treatments (cont)

Where?

2) estimate the causal effects of specific properties extracted from text:

e.g: how appealing to civic duty via mails impacts on voter turnout in Michigan 2006 primary election (Gerber et al 2008)

→ factors are latent properties of the text for which we need a measurement model.

CI with text as treatments (cont)

Goal? The causal relationship that (specific aspects of the) text has on downstream decisions, behaviors, and outcomes.

Challenges?

1) Ignorability is typically violated.

Why? People often choose the text they read for reasons related to outcome of interests. (e.g.: political affiliation, mood, education level, gender)

CI with text as treatments (cont)

Challenges?

2) Positivity is often violated.

Why? Suppose we want to know if politeness causally affect e-mail response times. Even if the text contains all confounders (e.g.: topic, tone, writing styles), a polite email is unlikely to contain a certain writing style (e.g., profane).

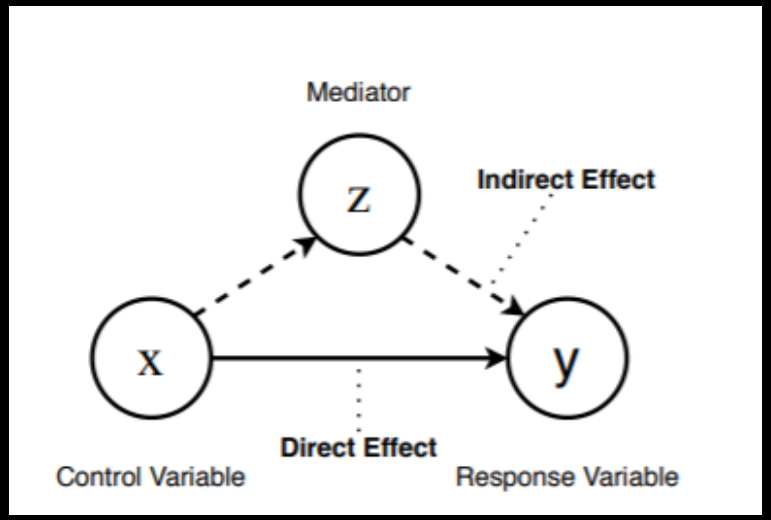
CI with text as mediator

Causal mediation analysis (Hicks and Tingley 2011 for Stata, Tingley et al., 2014 for R, Statsmodel in Python) introduces a mediator into the causal graph.

The mediator decouples the total effect into a direct and indirect effect (of intermediaries).

E.g: Vaccine causes headaches -> subjects take aspirin (mediators) -> impact on protection level

→ To understand the process in which the treatment causally affects the outcome.



CI with text as mediator: an application

How grammatical gender bias in large pre-trained Transformer-based language models. (Vig et al., 2020)

Major concern in word representations, both static word embeddings and contextualized word representations.

(You can check their Github code here: https://github.com/sebastianGehrmann/CausalMediationAnalysis)

Given a prompt u : The nurse said that…, a language model is asked to generate a continuation:

Stereoptypical candidate: SHE . Anti-stereotypical candidate: HE

→ A biased model assigns a higher likelihood to SHE than HE

$$p_\theta(SHE|u) > p_\theta(HE|u)$$

→ A measure of grammatical gender bias in the model:

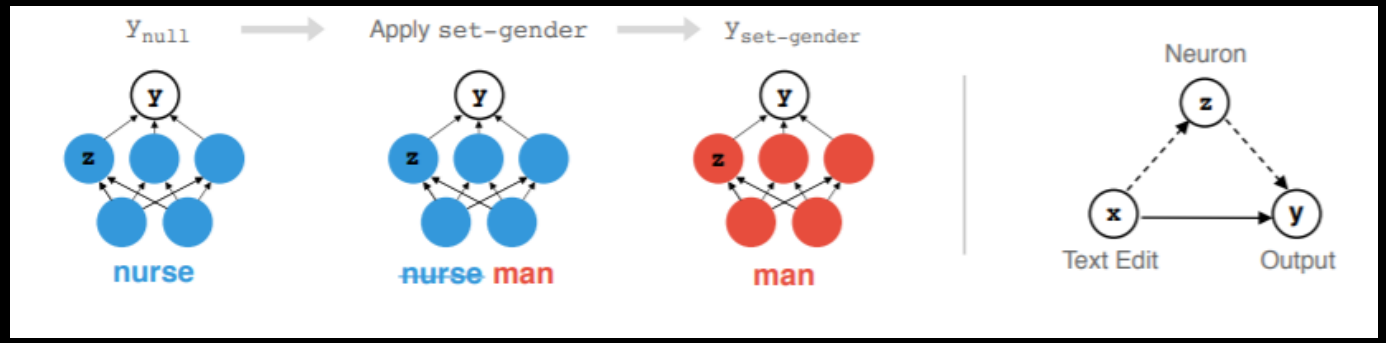
$$y(u) = \frac{p_\theta(HE|u)}{p_\theta(SHE|u)}$$

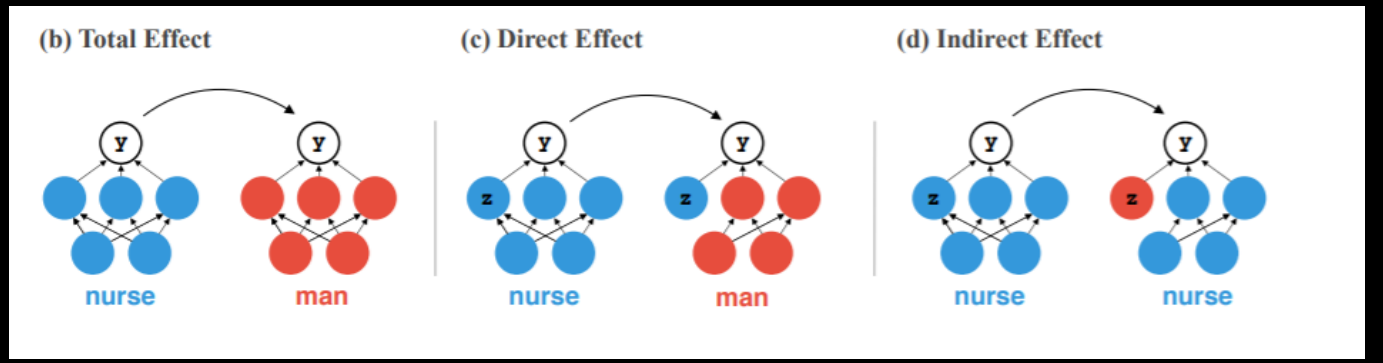
$y(u)$ >1 ?

$y(u)$ <1 ?

$y(u)$ =1 ?

CI with text as mediator: an application





Opportunities

NLP causal reasoning (data augmentation, distributional criteria)

→ read more in detail in Feder et al., 2021: Causal Inference in NLP

Econ ft. Machine Learning (EconML packages with advanced methods e.g. double machine learning, causal forests, deepiv, doubly robust learning, dynamic double machine learning)

Practice time ☺

Open your Google Colab

https://github.com/httn21uhh/Text-Analysis-for-Social-Sciences-in-Python

→ W13_causalML_Boston_housing.ipynb

- Download and put them in the same directory/environment. Open the task.

This week…

- Correlation vs. Causality & Causal Interpretation of Boston housing price data set
- Unmute yourself to ask questions at any point, including when you think things go too fast/slow for you.