Causal Inference in NLP

Estimation, Prediction, Interpretation and Beyond

2023-11-06

Abstract

- Fundamental goal of (social) scientific research: **explore causal relationship**
- However, causality is understudied in NLP, which focuses more on prediction
- This distinction is beginning to fade, this paper tries to accelerate this process by introducing:
  - statistical challenges of estimating causal effects with text (**text as outcome, treatment or confunders**)
  - use CI to improve the **robustness, fairness and Interpretation** of NLP models.

Causal Inference Overview

- Classical example: drug therapy on disease progression

. . .

- Counterfactual world cannot be observed

. . .

- Estimate casual effects with observational data face challenges

. . .

```
-   fortunately we have PSM,IV,RD,DID,RCT,etc.
```

. . .

- Applying CI into NLP data faces new fundamental challenges

NLP Overview

- Any correlation is admissible, regardless of the underlying causal relationship

. . .

- High-stakes scenarios
- data distribution in train and new dataset
- black-box

Exmaples CI $ NLP

1. Social media gender indication and post popularity

spurious correlation

. . .

- confounder topic (text as confounder)
- Gender signal effects on sentiment of the posts(text as outcome)
- Writing style effects on post likes(text as treatment)

. . .

2. Detect clinical diagnoses from the textual narratives

- frequency of the target clinical condition and the writing style of the narratives
- its prediction performance decreases with new data (overfitting, noise)

Causal Estimands

- ATE

$$ATE = E[Y(1) - Y(0)]$$

. . .

- CATE

$$CATE = E[Y(1) - Y(0)|G]$$

. . .

- Ignorability

$$T \perp Y(a) \quad \forall a \in \{0, 1\}$$

. . .

- Without RCT, confounders will lead to biased estimation of ATE
- Experience change gender icon write posts with high popularity

. . .

- Positivity

$$0 < Pr(T = 1|X = x) < 1, \forall x$$

. . .

- Consistency

$$T = a \Leftrightarrow Y(a) = Y, \forall a \in 0, 1$$

- versions of treatment $ interference

Previous Approaches

- Question: X Gender signal Y harassment Confounder(Z) Topic
- NLP $ PSM: extract topic from text and then conduct PSM
- The behind logic: identify the confonding property of text and control

. . .

- Caveat: Ignorability
- Requires domain expertise to illustrate and evaluate the consequence of violation

. . .

- Positivity can hardly hold
- (When some topics only appear on Female posts)

Text as Outcome

Suspension on toxicity decreasion

- Traditional methods: Annotate the post content with toxicity score
- NLP approach: extract toxicity from text by dimensional reduction
  - Consistency Assumption
  - If we use cluster model to classify post content
    - this cluster model trained on all users' text data The text data was influenced by the treatment

- Everyone' outcome is related to the treatment of others

- Solution: conduct measurement on samples and estimate effects on held-out data sample

Causal Effects with Textual Treatment

What makes a post offensive? (Second-preson pronouns)

- Previous approaches: "discover" the treatment by NLP
- Challenges
  - conditional ignorability(can we disentangle T from other aspect from text)
  - positivity
  - consistency

Future Work

Heterogeneous effects

- People read and interpret the text differently.
  - Random forests on tabular data
  - Opportunity: use NLP to finds the text features that captures subgroups where subgroup effect varies

Representation learning

- Extract latent aspects that satisfy:
  - positivity is satisfied
  - confound-ing information is not discarded
  - noisily-measured outcomes or treatments enable accurate causal effect estimate

Thanks

Resources of NLP in Social Science

Computational Analysis of Communication