Sensitivity Analysis for Causal Mediation through Text: an Application to Political Polarization

Huan Chen, Nico Hudaff and Julius Koch

17.05.2022

Outline

Motivation [15]

Political polarization is defined as the divergence of political attitudes away from the center. This leads to problems dealing with facts and values that conflict with the agenda of one's own faction. Electorates often lose faith in public institutions and support for democracy can decline. This highlights the importance of understanding the salient dynamics of polarization.

- We want to study the effect of conversations on political polarization
- We hope to identify what particular conversation features lead to depolarization

Research Questions [15]

- Relationship between political polarization and dialogue with persons from other end of political spectrum
  - Randomized experiment
  - Anonymous online chat
  - Assigned to discuss either immigration or gun control
  - Control group did not have a conversation
- Causal mediation: direct effect of simply having a conversation and the indirect effect of the conversation content

Data and Sources [1, 2, 15]

- Previous work - original research question: does having a conversation with a political opposite reduce political polarization
  - Downloaded chatting app with partners randomly and anonymously assigned
  - Topics were gun control and immigration
  - Surveys were conducted before and after the conversations
  - Politics were not mentioned in recruitment dialogue, and
  - Control group did not have conversation
- Origins of data: [1, 2]

Challenges [15]

We cannot rely on the randomization to identify what conversation features caused the depolarization

- The conversation text is properly thought of as a mediator variable, something causally affected by treatment that also affects the outcome
- Mediators are missing for control group, so we cannot compare treatment and control text

Solutions [15]

- Politeness as a mediator
  - prior work [3, 10, 5, 9, 7] identifies politeness as a key feature in producing persuasive text
  - measure politeness of messages received by individuals in treatment
  - link the measure to depolarization
- Mediation sensitivity
  - impute politeness measures for control group
  - preserve the relationship between politeness and depolarization among the treated
  - test how large the difference in politeness between treatment and control must be to observe significant mediation

Mediation Analysis [15]

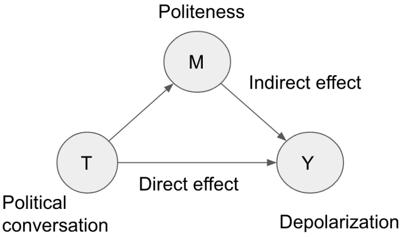

Figure: Casual Inference Framework

The total effect can be decomposed by:

$$TE := \underbrace{\mathbb{E}[Y(1, M(1)] - \mathbb{E}[Y(0), M(1)]}_{\text{direct effect}} + \underbrace{\mathbb{E}[Y(0), M(1)] - \mathbb{E}[Y(0), M(0)]}_{\text{indirect effect}}$$

Causal Inference with Text Data [15]

- difficulties are that text is outcome, mediator, confounder and treatment at the same time
- text as a confounder previous works [13, 12, 8]
- generally most of the methods try to compare treated and control units with similar text
- transforming high dimensional text data into low dimensional or binary representation

Politeness Measure [15]

- question: what is the mediator of the treatment of having a political conversation?
- one approach is the politeness of the conversation
- previous works can help to identify politeness [3, 10, 5, 9, 7]
- R package "politeness" based on [17] was implemented
- standardize the politeness measure for each feature to have mean of 0
- sum the politeness measure for each message received

Modelling Mediator and Outcome [15]

Outcome (Y) given treatment (T), mediator (M), measured covariates (X) and individual error term (ϵ):

$$Y_i = \tau T_i + \alpha M_i + \gamma X_i + \epsilon_i \tag{1}$$

Mediator (M) given treatment (T), measured covariates (X) and individual error term (ν):

$$M_i = \beta T_i + \theta X_i + \nu_i \tag{2}$$

Indirect and Direct Effect [15, 11]

$$Y_i = \tau T_i + \alpha M_i + \gamma X_i + \epsilon_i$$
$$M_i = \beta T_i + \theta X_i + \nu_i$$

- The direct effect ($\mathbb{E}[Y(1, M(1)) - Y(0, M(1))]$) is given by τ based on equation 1
- The indirect effect ($\mathbb{E}[Y(0, M(1)) - Y(0, M(0))]$) is given by $\alpha\beta$ based on equations 1 and 2

Therefore the total effect is given by $TE := \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \tau + \alpha\beta$

Problem estimating β [15]

$$M_i = \beta T_i + \theta X_i + \nu_i$$

- due to missing control group from original experiment β is not able to be estimated
- needs to estimate the β, i.e. estimate M by $\tilde{M}$
- what is relationship of politeness and having a (political) conversation?
- $\mathbb{P}(Y, M | T = 1, X)$ is given, this will help in estimating $\tilde{M}$

Simulating $\tilde{M}$ - Conditions [15]

$$Y_i = \tau T_i + \alpha M_i + \gamma X_i + \epsilon_i$$
$$M_i = \beta T_i + \theta X_i + \nu_i$$

- the total effect ($\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$) should remain constant with changing values for $\tilde{M}$, since it is estimated by Y and T
- α shouldn't change as well, due to describing the relationship between the outcome (depolarization) and mediator (politeness)
  - $\alpha\beta$ wouldn't describe the indirect effect anymore
- we can study decomposition of indirect and direct effects based on the change of β

Simulating $\tilde{M}$ - approaches [15]

$$Y_i = \tau T_i + \alpha M_i + \gamma X_i + \epsilon_i$$

- naive approach: random draws from M_i to get the $\tilde{M}$
  - will lead to $\alpha \to 0$ as $n \to \infty$, while n being the number of draws (Law of large numbers [14])
  - would lead to independence of Y and M
- estimate M based on X among treated
  - $\tilde{M} = X_c(X_t^T X_t)^{-1} X_t^T M_t$
  - simple regression (OLS)
  - subscript c indicating values in control group and subscript t values in treated group
  - one can prove that this regression will lead to:
    - $\hat{\alpha} = \alpha$
    - $\hat{\beta} = 0$
    - $\hat{\tau} = TE$

Simulating $\tilde{M}$ - setting β [15]

$$Y_i = \tau T_i + \alpha M_i + \gamma X_i + \epsilon_i$$
$$M_i = \beta T_i + \theta X_i + \nu_i$$

- using the OLS estimate for $\tilde{M}$
- choosing a $C \in [0, \frac{TE}{\alpha}]$ with $TE = \tau$
- setting:
  - $\beta = C$
  - $\tilde{M} = \tilde{M} - C$

Topic Modelling - Raw Text Method [15]

- alternative approach for modelling the mediator M: topic modelling
- two implemented methods:
  - LDA [16]
  - supervised Indian Buffet Process (sIBP) [6]
- according to [4, 12] these methods are commonly used
- pre-processing steps:
  - lower case
  - removing stop words
  - removing words with less than 1% appearance of the documents
- 819 documents with 3715 words
- grouping 4 to 12 topics

Polarization and Politeness Results [15]

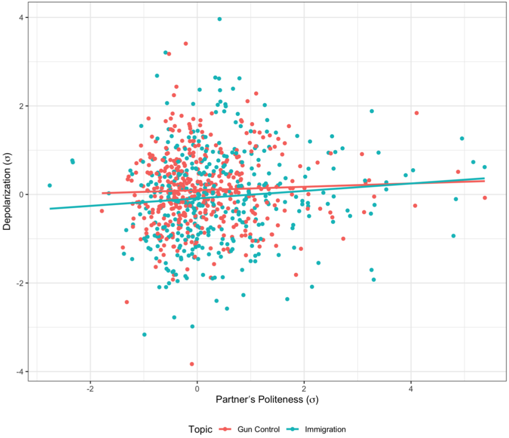

Figure: Depolarization and Politeness Graph

Polarization and Politeness Results cont. [15]

	All	Dem	Rep
	(1)	(2)	(3)
Politeness	0.069*	0.119*	0.002
	(0.035)	(0.046)	(0.052)
Constant	0.160	0.565	−0.294
	(0.215)	(0.288)	(0.312)
Observations	819	408	411

Figure: Politeness and Depolarization Table

Mediation Sensitivity Results [15]

	Treatment Only	Total Effect	Imputed Politeness
	(1)	(2)	(3)
Treatment		0.156*	0.156*
		(0.076)	(0.076)
Politeness	0.069*		0.069*
	(0.035)		(0.035)
Constant	0.159	0.119	0.134
	(0.214)	(0.202)	(0.202)
N	819	1,037	1,037

Figure: Politeness Imputation Results

Mediation Sensitivity Results cont. [15]

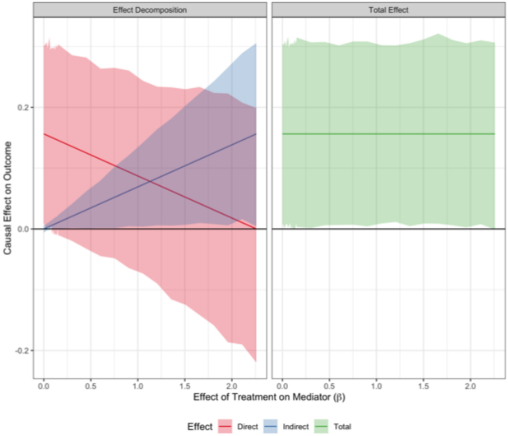

Figure: Decomposition of Treatment Effects by Simulated Effect on Mediator

Topic Modelling Results [15]

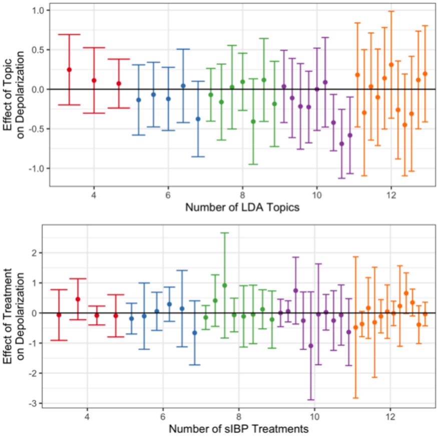

Figure: Topic Modelling Using 50/50 Training/Test Approach

Summary of Results [15]

- found relationship between politeness and depolarization
- high β implies politeness is correlated with having a political conversation (indirect effect)
- low β implies no correlation between politeness and having a political conversation (direct effect)
- topic modelling results were insignificant

Future Research

- do another experiment but have the control group have a non-political conversation (e.g. about sports or entertainment)
- conduct more research on mediation effects of text content on emotional responses with more data

Thank you for your attention

References I

[1] Christopher A. Bail et al. "Exposure to opposing views on social media can increase political polarization". In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221. DOI: 10.1073/pnas.1804840115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1804840115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1804840115.

[2] Aidan Combs et al. "Anonymous cross- party conversations can decrease political polariza- tion: A field experiment on a mobile chat platform". Unpublished Manuscript. 2021.

References II

[3] Cristian Danescu-Niculescu-Mizil et al. "A computational approach to politeness with application to social factors". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 250–259. URL: https://aclanthology.org/P13-1025.

[4] Naoki Egami et al. *How to Make Causal Inferences Using Texts*. 2018. DOI: 10.48550/ARXIV.1802.02163. URL: https://arxiv.org/abs/1802.02163.

References III

[5] Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. "Incorporating Politeness across Languages in Customer Care Responses: Towards building a Multi-lingual Empathetic Dialogue Agent". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4172–4182. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.514.

[6] Christian Fong and Justin Grimmer. "Discovery of Treatments from Text Corpora". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1600–1609. DOI: 10.18653/v1/P16-1151. URL: https://aclanthology.org/P16-1151.

References IV

[7] Dongyeop Kang and Eduard Hovy. *Style is NOT a single variable: Case Studies for Cross-Style Language Understanding*. 2019. DOI: 10.48550/ARXIV.1911.03663. URL: https://arxiv.org/abs/1911.03663.

[8] Katherine Keith, David Jensen, and Brendan O'Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5332–5344. DOI: 10.18653/v1/2020.acl-main.474. URL: https://aclanthology.org/2020.acl-main.474.

References V

[9] Aman Madaan et al. “Politeness Transfer: A Tag and Generate Approach”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1869–1881. DOI: 10.18653/v1/2020.acl-main.169. URL: https://aclanthology.org/2020.acl-main.169.

[10] Tong Niu and Mohit Bansal. “Polite Dialogue Generation Without Parallel Data”. In: *Transactions of the Association for Computational Linguistics* 6 (July 2018), pp. 373–389. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00027. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00027/1567654/tacl_a_00027.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00027.

References VI

[11] Judea Pearl. "Direct and Indirect Effects". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI'01. Seattle, Washington: Morgan Kaufmann Publishers Inc., 2001, pp. 411–420. ISBN: 1558608001.

[12] Margaret Roberts, Brandon Stewart, and Richard Nielsen. "Adjusting for Confounding with Text Matching". In: *American Journal of Political Science* 64 (July 2020). DOI: 10.1111/ajps.12526.

References VII

[13] Amrita Saha et al. "Complex Program Induction for Querying Knowledge Bases in the Absence of Gold Programs". In: *Transactions of the Association for Computational Linguistics* 7 (Apr. 2019), pp. 185–200. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00262. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00262/1924356/tacl_a_00262.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00262.

[14] Kelly E. Sedor. "The Law of Large Numbers and its Applications". In: 2015.

References VIII

[15] Graham Tierney and Alexander Volfovsky. "Sensitivity Analysis for Causal Mediation through Text: an Application to Political Polarization". In: *Proceedings of the First Workshop on Causal Inference and NLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 61–73. DOI: 10.18653/v1/2021.cinlp-1.5. URL: https://aclanthology.org/2021.cinlp-1.5.

[16] Victor Veitch, Dhanya Sridhar, and David Blei. "Adapting Text Embeddings for Causal Inference". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by Jonas Peters and David Sontag. Vol. 124. Proceedings of Machine Learning Research. PMLR, Aug. 2020, pp. 919–928. URL: https://proceedings.mlr.press/v124/veitch20a.html.

References IX

[17] Michael Yeomans, Alejandro Kantor, and Dustin Tingley. "The politeness Package: Detecting Politeness in Natural Language". In: *The R Journal* 10.2 (2018), pp. 489–502. DOI: 10.32614/RJ-2018-079. URL: https://doi.org/10.32614/RJ-2018-079.