Estimating Causal Effects of Tone in Online Debates

Dhanya Sridhar[1] , **Lise Getoor**[2]
[1]Columbia University
[2]UC Santa Cruz
ds3778@columbia.edu

Abstract

Statistical methods applied to social media posts shed light on the dynamics of online dialogue. For example, users' wording choices predict their persuasiveness [Tan *et al.*, 2016; Danescu-Niculescu-Mizil *et al.*, 2012a] and users adopt the language patterns of other dialogue participants [Danescu-Niculescu-Mizil *et al.*, 2011; 2012b]. In this paper, we estimate the causal effect of reply tones in debates on linguistic and sentiment changes in subsequent responses. The challenge for this estimation is that a reply's tone and subsequent responses are confounded by the users' ideologies on the debate topic and their emotions. To overcome this challenge, we learn representations of ideology using generative models of text. We study debates from 4FORUMS.COM and compare annotated tones of replying such as emotional versus factual, or reasonable versus attacking. We show that our latent confounder representation reduces bias in ATE estimation. Our results suggest that factual and asserting tones affect dialogue and provide a methodology for estimating causal effects from text.

1 Introduction

Debates on online forums or social media sites provide observational data for studying discourse. Current understanding draws upon theories such as linguistic accommodation, which states that dialogue participants change and vary their wording styles to mirror one another [Gallois and Giles, 2015; Giles and Baker, 2008]. Statistical methods applied to social media have shown evidence of linguistic style accommodation [Danescu-Niculescu-Mizil *et al.*, 2011], power dynamics [Danescu-Niculescu-Mizil *et al.*, 2012b] and varying persuasiveness of argumentation styles [Tan *et al.*, 2016].

In this paper, we focus on online debates. We ask the causal question of how the tone used to reply in a debate affects changes in linguistic style and sentiment. To illustrate the setting, consider a snippet of a debate between two users, A and B, on a given topic. User A posts her opinion on the topic to which user B replies with a nasty tone. User A writes a second post, responding to B's post. The goal is to examine the change in A's sentiment or linguistic style between her first and second post. For example, we may observe that between her first post and her second post in response to B, A's negative sentiment increased. We study how A's sentiment might have changed had B been nice instead of nasty in the reply. We consider such sequences of three posts within debates and cast the tone of the first reply as the treatment. Formally, we estimate the average treatment effect of reply tone on changes in sentiment and linguistic style.

The challenge for this estimation is that the ideologies encoded in A and B's posts, and A's initial sentiment affect both B's reply tone and A's subsequent response. For example, consider a debate between A and B on gun control. Examples of opposing ideologies that influence the debate are strong opposition to gun violence versus strict interpretations of the constitution. Ideological differences or innate negativity from A provoke both B's nasty tone and A's subsequent reactions. Valid causal inference requires adjusting for these confounders when estimating the treatment effect. While sentiment analysis tools are available for extracting posts' sentiment, modeling the latent ideologies that underpin a particular debate requires careful consideration. This paper proposes representations of ideologies learned from debates to adjust for confounding.

In recent social media analyses, adjusting for attributes such as discussion topic, post authors, timing of posts and posting frequency has been useful to understand post likeability, antisocial behavior and emoji use [Tan *et al.*, 2014; Jaech *et al.*, 2015; Cheng *et al.*, 2015; Pavalanathan and Eisenstein, 2015]. The adjustments are typically performed by only comparing posts that have similar values of the confounder. Our approach requires adjusting for the underlying facets of a debate, an unobserved and multi-dimensional confounder. We use a generative model of text to learn latent representations of posts to this end.

Main Idea. The goal of this paper is to estimate the causal effects of tones used in debate replies on other users' change in linguistic style and sentiment. We identify treatments (tone) and outcome representations which capture the change in sentiment and linguistic style across sequences of posts. We use three plug-in estimators of average treatment effects: regression, inverse propensity weighting (IPW), and augmented IPW. To adjust for confounding in these estimates, we find latent representations of posts that capture the underlying ideological viewpoints of the debate. Our contributions

include:

- Formulating the problem of estimating the effects of tone on subsequent dialogue within the framework of causal inference.
- Learning latent representations of ideologies in debates from generative language models to represent confounders.
- Validating the consistency of estimated effects using three different estimators and examining multiple tones of reply in online debates.

We study 4FORUMS.COM, an online debate forum corpus that includes annotations for multiple reply tones including nasty versus nice, or emotional versus factual [Walker *et al.*, 2012b]. Through comparisons against a naive confounder representation and studies across reply tones, we examine the implications of various modeling choices. With these findings, we highlight guidelines for estimating treatment effects using text from social media. We also find that factual replies significantly affect how users' vary their linguistic style and sentiment between posts.

2 Related Work

Prior work on online debate forums primarily focus on supervised prediction tasks. Debate text and reply structure between users has been used to predict stance, sentiment and reply polarity [Abbott *et al.*, 2011; Walker *et al.*, 2012a; Hasan and Ng, 2013; Sridhar *et al.*, 2015; Misra and Walker, 2013; Rosenthal and McKeown, 2015]. Related work has also used the Change My View forum on Reddit.com to predict persuasiveness from styles of argumentation and characterize logical fallacies [Tan *et al.*, 2016; Wei *et al.*, 2016; Habernal *et al.*, 2018].

Similarly, unsupervised methods have been applied to analyze dialogue. Statistical models have been proposed to quantify linguistic accommodation both on Twitter and U.S. Supreme Court arguments [Danescu-Niculescu-Mizil *et al.*, 2011; 2012b]. In contrast, we formulate an approach based on causal inference.

Existing work on applying causal inference methods to social media focuses on controlling for confounding, or inferring treatments and outcomes from text. One line of work controls for observable confounders such as topic [Tan *et al.*, 2014; Jaech *et al.*, 2015], timing of posts [Jaech *et al.*, 2015] and the post author [Tan *et al.*, 2014]. Another line of research uses social media posts to estimate the effect of exercise on mood by inferring both exercise habits and mood from text [Dos Reis and Culotta, 2015; Olteanu *et al.*, 2017]. In a different line of work, embeddings of text have been used as proxy confounders to study causal effects on paper acceptance [Veitch *et al.*, 2019].

3 Technical Background

We review estimating the average treatment effect (ATE) for binary treatments from observational data. We have n iid observations called units, $i = 1 \dots n$. Each unit is treated or not, and we denote this treatment assigment $T_i \in \{0, 1\}$. We say $Y_i(1)$ is the potential outcome if we treat unit i (set $T_i = 1$), and analogously, $Y_i(0)$ if we do not treat i (set $T_i = 0$). The average treatment effect (ATE) compares potential outcomes:

$$ATE = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \tag{1}$$

However, we only observe one outcome for each unit, conditioned on its assigned treatment $Y_i(T_i)|T_i$. If we compute the ATE above by simply averaging over treated and untreated populations, the estimate will typically be biased because $Y(0), Y(1)$ are not independent of the assigned treatments T. Put simply, knowing the treated and untreated units gives us information about their outcomes.

In observational studies, this bias occurs because variables Z, called confounders, may affect both the treatment and outcome. If we observe the confounders for each unit, Z_i, then we have $Y(0), Y(1) \perp T|Z$, the condition called ignorability. In this case, the ATE, which we denote ψ, is identifiable as a parameter of the observational distribution by a theorem called adjustment:

$$\psi = \mathbb{E}_Z\big[\mathbb{E}_{Y|T=1}[Y|Z, T=1] - \mathbb{E}_{Y|T=0}[Y|Z, T=0]\big] \tag{2}$$

In plain English, the ATE is: how do treated and control units differ in outcome when we average over the varying rates at which units receive treatment? We refer to work by Pearl and Rubin for an in-depth treatment of causal inference [Pearl, 2009; Rubin, 2005].

3.1 Estimators for ATE

Drawing from extensive work on estimating ATEs, we present three estimators for ψ. We will return to using these estimators in our empirical study. The first estimator fits expected outcomes $Q(Z, T) = \mathbb{E}[Y|Z, T]$ from the observations, e.g., with linear regression. The corresponding ATE estimate, $\hat{\psi}^{\text{MLE}}$ is:

$$\hat{\psi}^{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} Q(Z_i, 1) - Q(Z_i, 0) \tag{3}$$

The second estimator reweights observed outcomes using the propensity score, $P(T = 1|Z)$. The resulting inverse propensity weighting (IPW) estimator is:

$$\hat{\psi}^{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i T_i}{P(T_i = 1|Z_i)} - \frac{Y_i(1 - T_i)}{1 - P(T_i = 1|Z_i)} \tag{4}$$

The final estimator, augmented-IPW (AIPW), interpolates between the two estimators [Robins *et al.*, 1994; Van der Laan and Rose, 2011]. It has been shown that the AIPW estimator satisfies double robustness: it retrieves consistent estimates if either the propensity score or outcome model is correct even if the other is misspecified. It is:

$$\hat{\psi}^{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i(Y_i - Q(Z_i, 1))}{P(T_i = 1|Z_i)} - \frac{1 - T_i(Y_i - Q(Z_i, 0))}{1 - P(T_i = 1|Z_i)} + Q(Z_i, 1) - Q(Z_i, 0) \tag{5}$$

All three estimators rely on measurements of confounders Z. We will see that in debate threads, we must recover the confounders from high-dimensional text. A key idea of this paper is to use text data to find a representation for Z.

4 Dataset

To estimate the ATE of tone, we use the 4FORUMS.COM corpus collected and annotated as part of the Internet Argument Corpus [Walker *et al.*, 2012b]. 4FORUMS.COM has been used to predict users' stances on topics, disagreements between users, and sarcasm use [Lukin and Walker, 2013; Walker *et al.*, 2012a; 2012c; Sridhar *et al.*, 2015].

4FORUMS.COM is a collection of debate discussions, each belonging to a topic such as "evolution" or "climate change." For some pairs of posts called **quote-response** pairs, 4FORUMS.COM includes annotations about the reply, obtained using Amazon Mechnical Turk. A quote-response pair is a post and its reply where the replier quotes the original poster and responds directly to the quoted statement. The response is annotated by multiple annotators along four dimensions which we refer to as **reply types**: nasty/nice, attacking/reasonable, emotional/factual, questioning/asserting. Each reply type has two opposing polarities (e.g., nasty or nice) which we refer to as its **tone**. The annotation score for each type ranges from -5 to 5, where negative values correspond to the antagonistic tone such as nasty or attack and positive values map to tones such as reasonable or factual.

We select the four debate topics with the most quote-response annotations: "abortion", "gay marriage", "evolution", and "gun control". Each debate topic has on roughly 1200 quote-response annotations. Following prior work, we use the mean score across annotators and discard annotations with a mean score between -1 and 1 [Misra and Walker, 2013; Lukin and Walker, 2013; Sridhar *et al.*, 2015]. In the next section, we formalize the use of these annotations as treatments to estimate causal effects.

5 Problem Statement

To study causal effects in debate threads, we first introduce **post triples**. A **post triple** $t_i = (p_i^1, p_i^2, p_i^3)$ is an ordered sequence of three posts where each post p_i^j belongs to the i-th triple and appears j-th in the sequence. The author of post p_i^j is denoted by a_i^j. The triples we consider have the property that $a_i^1 = a_i^3$, i.e., the same user authors the first and last posts. Based on the discussion in which the triple appears, the triple t_i has a debate topic τ_i. We will refer to p_i^1 as the original post and to p_i^2 as the reply post.

Each triple we study has a quote-response annotation for p_i^2 towards p_i^1. Given the reply type α of the annotations and its mean score, we binarize the values by considering those ≤ -1 as 0 and ≥ 1 as 1. Replies are thus converted to binary negative or positive tone, such as nasty or nice. For each triple t_i and reply type α, the tone of reply post p_i^2 toward p_i^1 gives the treatment assignment $T_i = R_i^\alpha \in \{0, 1\}$ for the triple.

Outcomes. The next problem is to quantify the outcome of interest: changes between p_i^3 and p_i^1 after receiving reply p_i^2. We rely on the Linguistic Inquiry and Word Count (LIWC) tool [Pennebaker *et al.*, 2007]. LIWC is a dictionary which maps an extensive set of English words to categories that capture both lexical and semantic choices. Several text classification and statistical analysis tasks have represented posts with counts of LIWC categories [Anand *et al.*, 2011; Abbott *et al.*, 2011; Walker *et al.*, 2012a; Danescu-Niculescu-Mizil *et al.*, 2011].

We first combine LIWC categories into groups which we call **category types** that measure positive sentiment, negative sentiment, and linguistic style. Fig. 1 shows each category type. For the sentiment groupings, we select the categories related to positive and negative emotion as listed on the LIWC website. For linguistic style, we use the categories identified in prior work for linguistic style accommodation [Danescu-Niculescu-Mizil *et al.*, 2011].

Given a category type, the frequency of words in p_i^j belonging to each category gives a vector representation of the post. We can construct such vector representations for p_i^1 and p_i^3. Formally, for a category type c and reply type α, the outcome $Y_i^{c,\alpha}$ for triple t_i is the Euclidean distance between the vector representations for p_i^1 and p_i^3. This strategy suggests many possible vector representations of posts including word embeddings [Mikolov *et al.*, 2013].

We state the ATE estimatation problem for these debate triples. For all configurations of c and α, we estimate:

$$\psi = \mathbb{E}\big[\mathbb{E}[Y^{c,\alpha}|Z, R^\alpha = 1] - \mathbb{E}[Y^{c,\alpha}|Z, R^\alpha = 0]\big]$$

This estimates the mean difference in text changes between users receiving a positive-tone reply and those receiving negative-tone ones. The main challenge is to find a representation for Z that captures plausible confounders in debates.

6 Constructing Confounder Representations

In a post triple of interest, the debate topic, latent ideologies of each author within the topic, and sentiment of the original author are confounders. That is, these variables plausibly influence both the treatment (reply tone) and the outcome (change between posts). Prior work has shown that text in political debates can be mapped into a lower dimensional space that corresponds to the moral or ideological facets of that debate topic [Johnson and Goldwasser, 2018; Misra *et al.*, 2015; Iyyer *et al.*, 2014; Boydstun *et al.*, 2013]. Unsupervised approaches have been used to discover word-clusters that correspond to these frames directly from text [Iyyer *et al.*, 2014]. Here, we fit an unsupervised generative model of text to learn ideology representations.

Ideology Representation. We use the latent Dirichlet allocation (LDA) topic model [Blei *et al.*, 2003] to recover unobserved ideologies. The observations w_{ij} are counts of word j in document i. The generative process is:

$$\beta_k \sim \text{Dirichlet}(\gamma) \quad (6)$$
$$\theta_i \sim \text{Dirichlet}(\alpha) \quad (7)$$
$$z_{ij} \sim \text{Multinomial}(\theta_i) \quad (8)$$
$$w_{ij} \sim \text{Multinomial}(\beta_{z_{ij}}) \quad (9)$$

Each of the k latent topics β_k is a distribution over the words in the vocabulary. Each document-level latent variable θ_i is a distribution over topics. For a document i, each word w_{ij} is drawn by sampling a topic from the document's distribution over topics and then sampling a word from that topic.

The posterior expected values $\mathbb{E}[\theta_i]$ and $\mathbb{E}[\beta_j]$ converge to the optimal values of θ_i and β_j. This convergence property

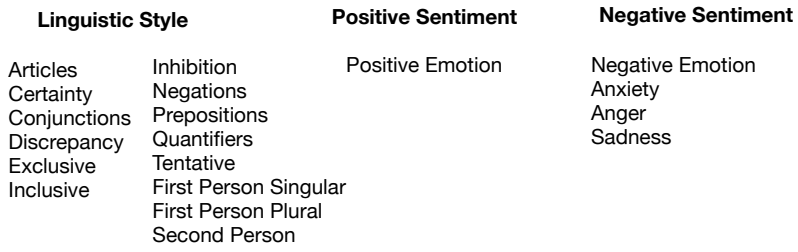

Figure 1: We group LIWC categories into three types: linguistic style, positive sentiment, negative sentiment. We use the groupings for vector representations of posts which we can use to measure the outcomes of interest: changes between the first and last post of a triple.

allows us to substitute $\mathbb{E}[\theta_i]$ as a confounder for each post. LDA is fit using variational inference.

We fit LDA for each debate topic τ with the observed word counts across posts from that topic. By conditioning on τ, the confounder representation incorporates both the debate topic and finer-grained ideology. The inferred mean proportions over latent topics $\mathbb{E}[\theta_i]$ is the embedding for each post. For each triple t_i, we concatenate the embeddings for posts p_i^1 and p_i^2 to include in the confounder Z_i. Since the embeddings aim to approximate ideologies, including both p_i^1 and p_i^2 embeddings helps to further deconfound the effect of users' opposing or similar views on tone and word change.

Sentiment Representation. To represent the sentiment encoded in p_i^1, we use the same LIWC category types as we do for the outcomes. As before, we compute the frequency of words in p_i^1 that belong to each category. This gives us a vector representation of sentiment which we include in Z_i.

7 Empirical Results

The difficulty in validating causal effects, particularly in debates, is that there is no ground truth. Typically, causal estimation procedures are validated using simulated data but for text, realistic generative models do not exist. Thus, one of the paper's contributions is developing an evaluation strategy for text-based causal inferences. Our approach is three-pronged: 1) we assess the predictive performance of the key ingredients for estimation, the propensity score and expected outcome models; 2) we manually inspect the latent ideological topic ; 3) we compare causal effects across multiple estimators and against a naive confounder representation.

We found that: 1) causal estimation using our confounder representation reduces bias in the ATE estimates compared to using a naive confounder; 2) if we had not compared multiple estimators and instead used a single estimator like the high-variance IPTW (a common practice), we would have incorrectly reported effects; 3) the estimates suggest that emotional/factual and questioning/asserting tones elicit changes in linguistic style and emotion while nice/nasty or reasonable/attacking tones show no effect. Code and data to reproduce all results are available. [1]

Methods and Metrics. Using the latent confounder representation proposed, the goal is to estimate the ATE ψ of reply tone R_α on outcome $Y^{c,\alpha}$. In the empirical results below, we estimate ψ using three estimators ($\hat{\psi}^{\text{MLE}}, \hat{\psi}^{\text{IPW}}, \hat{\psi}^{\text{AIPW}}$) for all configurations of LIWC category type c and α that yield different treatments and outcomes. We report the unadjusted estimate, $\big[\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$, which will be biased. We compare against a naive representation, Z - Debate Topics Only, which only uses the debate topic τ_i without finer-grained ideologies.

We fit the propensity score $P(T = 1|Z)$ (used by $\hat{\psi}^{\text{IPW}}, \hat{\psi}^{\text{AIPW}}$) with logistic regression using the observed treatments and constructed confounder representations. We fit the expected outcomes, $Q(Z, T = 0), Q(Z, T = 1)$ (used by $\hat{\psi}^{\text{MLE}}, \hat{\psi}^{\text{AIPW}}$) with linear regression.

Experimental Setup. Besides the processing of quote-response pair annotations described in the Data section, we prepare the posts to fit LDA. We obtain all unigram tokens after lemmatizing and removing stop words from posts across all discussions for a given topic. We retain only those tokens which occur in more than 2% but in fewer than 80% of posts. For each topic, this yields a document-term-frequency matrix of roughly 30, 000 posts and 400 remaining terms after preprocessing. We fit LDA with $k = 50$ topics. For each reply type, the ATE is averaged over roughly 1500 triples. For the AIPW estimator $\hat{\psi}^{\text{AIPW}}$, we use a variant proposed to improve finite sample performance [Van der Laan and Rose, 2011].

Performance of Outcome and Propensity Models. The first step to validating ATEs is to verify that the models used in various estimators fit the observations well. Table 1 gives the root mean squared error (RMSE) and F1 for the expected outcomes $Q(Z, T)$ and propensity scores $P(T = 1|Z)$, respectively. We perform five-fold cross-validation on the triples. We report performances for all reply types α but for the outcome model which depends also on the LIWC category type, we show scores for positive sentiment outcomes for conciseness. Our code will reproduce the other RMSEs.

For the positive sentiment expected outcomes, our confounder Z-Full predicts with lower RMSE than Z-Debate Topics Only. The F1 score is also slightly improved by using Z-Full over Z-Debate Topics Only when predicting emotional/factual. While this is reassuring, for causal inference, unbiased estimation is more important than predictive performance. Below, we investigate causal estimation, where Z-Debate Topics Only shows undesirable consequences.

[1] github.com/dsridhar91/debate-causal-effects

Performance of Outcome (Pos. Sent, RMSE) and Propensity Models (F1)

Reply type	Z - Debate Topics Only			Z - Full		
	$Q(Z, T=1)$	$Q(Z, T=0)$	$P(T=1\|Z)$	$Q(Z, T=1)$	$Q(Z, T=0)$	$P(T=1\|Z)$
Nasty/Nice	3.5	2.7	0.89	2.6	2.4	0.89
Attacking/Reasonable	3.6	2.7	0.81	3.0	2.7	0.81
Emotional/Factual	2.4	5.1	0.69	2.2	5.1	0.72
Questioning/Asserting	3.1	4.4	0.80	2.9	4.1	0.79

Table 1: The expected outcome models (shown here for positive sentiment) using Z-Full generally improves over using Z-Debate Topic Only. The propensity score model performs comparably in both cases. We evaluate the models using 5-fold cross-validation. For expected outcomes, we report RMSE and F1 for the propensity score.

Top 10 words per latent topic (two shown per debate topic)

Abortion	Argument	Point	Zygote	Science	Human	Scientific	Argue	Becomes	Alive	Individual
Abortion	God	Bible	Faith	Love	Believe	Created	One	Everything	Free	Made
Gay Marriage	Majority	Black	Love	Right	Marry	Minority	White	Loving	Vote	Race
Gay Marriage	Moral	Faith	Morality	Animal	Good	Meant	Hurt	Wrong	One	Say
Evolution	Mutation	Selection	Gene	Change	Organism	Natural	Genetic	Random	DNA	Trait
Evolution	Bible	Christian	Genesis	Creation	Christianity	Creator	Literals	God	Word	Biblical
Gun Control	Government	Right	Constitution	State	Power	Law	Court	Supreme	Constitutional	Liberty
Gun Control	Violence	Time	Show	Record	Book	Likely	Incident	Stupid	Measure	Reported

Figure 2: Top words across latent topics from each debate topic chosen to illustrate ideologies captured by LDA. The latent topics are suggestive of viewpoints like morality and faith versus science and evidence.

ATE (and Standard Error) for Nasty/Nice

Estimator	Z - Debate Topics Only			Z - Full		
	Pos.	**Neg.**	**Ling.**	**Pos.**	**Neg.**	**Ling.**
Unadjusted	0.0	-0.8	-0.6	0.0	-0.8	-0.6
$\hat{\psi}^{\text{MLE}}$	0.0 (0.0)	-0.8 (0.1)	-0.6 (0.1)	0.0 (0.1)	-0.3 (0.1)	-0.3 (0.1)
$\hat{\psi}^{\text{IPW}}$	0.0 (0.2)	-0.7 (0.35)	-0.6 (1.0)	0.0 (0.3)	-0.2 (0.3)	-0.1 (1.0)
$\hat{\psi}^{\text{AIPW}}$	0.0 (0.2)	-0.8 (0.3)	-0.6 (0.5)	0.0 (0.1)	-0.3 (0.2)	-0.3 (0.4)

Table 2: The debate topics-only approach can overestimate treatment effects; it remains more biased (compared to the unadjusted estimate) than using Z-Full. We report ATE (and standard error) for Nasty/Nice reply type.

Latent Ideologies. Even if the ideology representations inferred using LDA are useful for predictive performance above, we carefully inspect the latent topics found by LDA. Since we assumed that ideology is a confounder, we want the latent topics to approximate ideologies. We inspected the top ten words associated with each latent topic across all debate topics. Fig. 2 shows these words for two latent topics from every debate topic as illustrative examples of the ideological views found. For example, in gun control debates, LDA finds topics that align with constitutional rights to bear arms and in evolution debates, there are contrasting topics that align with creationist versus scientific views. Our code includes simple visualization to inspect all latent topics.

ATE Estimation. We use Z-Debate Topic Only and Z-Full and apply the three estimators $\hat{\psi}^{\text{MLE}}, \hat{\psi}^{\text{IPW}}$ and $\hat{\psi}^{\text{AIPW}}$. We compare these estimates against the unadjusted estimate. Table 2 shows the ATEs (and standard error) for the nasty/nice reply type. When confounders are missing from adjustment, we expect the estimate to be closer to the biased, unadjusted estimate. Indeed, the results show that using Z - Debate Topics Only, omitting sentiment and ideology, consistently yields

ATE (and Standard Error) for Remaining Reply Types

Estimator	Attacking/Reasonable			Emotional/Factual			Questioning/Asserting		
	Pos.	**Neg.**	**Ling.**	**Pos.**	**Neg.**	**Ling.**	**Pos.**	**Neg.**	**Ling.**
Unadjusted	-0.1	-0.6	-0.6	-1.0	-0.8	-2.3	-0.5	-0.2	-1.7
$\hat{\psi}^{\text{MLE}}$	0.1 (0.1)	-0.2 (0.1)	-0.2 (0.1)	**-0.6 (0.1)**	**-0.3 (0.1)**	**-1.4 (0.1)**	**-0.3 (0.1)**	-0.2 (0.1)	**-1.2 (0.1)**
$\hat{\psi}^{\text{IPW}}$	0.1 (0.2)	-0.1 (0.3)	-0.4 (0.4)	-0.4 (0.3)	-0.2 (0.2)	-0.7 (0.8)	-0.3 (0.3)	-0.2 (0.2)	-1.1 (0.8)
$\hat{\psi}^{\text{AIPW}}$	-0.1 (0.1)	-0.2 (0.9)	-0.4 (0.4)	**-0.6 (0.2)**	**-0.3 (0.1)**	**-1.2 (0.3)**	-0.3 (0.2)	-0.2 (0.1)	**-1.2 (0.3)**

Table 3: Factual and asserting tones result in the first dialogue participant significantly decreasing changes in linguistic style. Factual tones may provoke decreased change in sentiment. We report ATE (and standard error) for all remaining reply types. Bolded numbers indicate that the ATE is significantly greater than zero.

estimates which are closer to the unadjusted estimate than using Z - Full. This is a key finding: estimation bias is reduced with the finer-grained confounder representation. After adjusting for Z - Full, the effects on dialogue outcomes from nasty/nice tones are not significant.

In Table 3, we proceed with our confounder representation Z - Full and study the remaining reply types: attacking/reasonable, emotional/factual, questioning/asserting. The $\hat{\psi}^{\text{MLE}}$ and $\hat{\psi}^{\text{AIPW}}$ estimators find significant effects, particularly for factual and authoritative tones. The $\hat{\psi}^{\text{IPW}}$ estimator yields similar ATE estimates but suffers from high variance. This is another key finding that comparing multiple estimators provides a form of validation: the propensity score-based $\hat{\psi}^{\text{IPW}}$ estimator is known to have high variance, and without the two remaining estimators, we may have concluded that no significant effects occur.

The $\hat{\psi}^{\text{MLE}}$ and $\hat{\psi}^{\text{AIPW}}$ estimators suggest that factual and asserting tones cause decreased changes in linguistic style: users' second posts remain closer to their original posts on average across triples. Further, these estimators suggest that both positive and negative sentiment changes are decreased when the tone is factual. The results suggest that users change their overall sentiment less when responding to factual arguments instead of emotionally charged ones. However, users may also maintain their original linguistic styles more when responding to factual or asserting tones. This finding on the role of factual and asserting tones may point to in-depth followup studies on persuasion and argumentation in debates.

Finally, the empirical studies reveal unexpected findings about causal estimation of treatments effects in debates. Interestingly, the choice of outcome representation matters: Table 3 in particular shows that changes to linguistic style are affected more than sentiment changes. A single outcome representation which had concatenated all LIWC categories or used word embeddings may have yielded different results.

8 Discussion

We study treatment effects in debates by estimating unobserved confounders from sequences of posts. We examine these interpretable embeddings in debates to find that they match known ideological views. The exercise of estimating treatment effects yields results of two flavors: 1) evidence that factual replies cause decreased change in linguistic style and sentiment, and 2) guidelines for practioners to estimate treatment effects from social media text.

We highlight areas of future study. In this work, we focus on ideology and sentiment as confounders. It is interesting to consider possible confounding from posts' timing and position in a discussion thread. A fruitful area of research is learning deep confounder (and outcome) representations for text while maintaining model interpretability, which we show in this paper is important for validating findings. Finally, validating causal effects in questions of social science research remains an open problem. Simulating outcomes from text is a line of future work.

Acknowledgements

This work is supported by NSF grants CCF-1740850 and IIS-1703331.

References

[Abbott *et al.*, 2011] Rob Abbott, Marilyn Walker, Jean E. Fox Tree, Pranav Anand, Robeson Bowmani, and Joseph King. How can you say such things?!?: Recognizing disagreement in informal political argument. In *ACL Workshop on Language and Social Media*, 2011.

[Anand *et al.*, 2011] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *ACL*, 2011.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(1), 2003.

[Boydstun *et al.*, 2013] Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop*, 2013.

[Cheng *et al.*, 2015] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. In *ICWSM*, 2015.

[Danescu-Niculescu-Mizil *et al.*, 2011] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words!: linguistic style accommodation in social media. In *WWW*, 2011.

[Danescu-Niculescu-Mizil *et al.*, 2012a] Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. You had me at hello: How phrasing affects memorability. In *ACL*, 2012.

[Danescu-Niculescu-Mizil *et al.*, 2012b] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *WWW*, 2012.

[Dos Reis and Culotta, 2015] Virgile Landeiro Dos Reis and Aron Culotta. Using matched samples to estimate the effects of exercise on mental health from twitter. In *AAAI*, 2015.

[Gallois and Giles, 2015] Cindy Gallois and Howard Giles. Communication accommodation theory. *The International Encyclopedia of Language and Social Interaction*, 2015.

[Giles and Baker, 2008] Howard Giles and Susan C Baker. Communication accommodation theory. *The International Encyclopedia of Language and Social Interaction*, 2008.

[Habernal *et al.*, 2018] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *NAACL*, 2018.

[Hasan and Ng, 2013] Kazi Saidul Hasan and Vincent Ng. Extra-linguistic constraints on stance recognition in ideological debates. In *ACL*, 2013.

[Iyyer *et al.*, 2014] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *ACL*, 2014.

[Jaech *et al.*, 2015] Aaron Jaech, Vicky Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. Talking to the crowd: What do people react to in online discussions? *Politics*, 7, 2015.

[Johnson and Goldwasser, 2018] Kristen Johnson and Dan Goldwasser. Classification of moral foundations in microblog political discourse. In *ACL*, 2018.

[Lukin and Walker, 2013] Stephanie Lukin and Marilyn Walker. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *NAACL*, 2013.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Misra and Walker, 2013] Amita Misra and Marilyn A Walker. Topic independent identification of agreement and disagreement in social media dialogue. In *SIGDIAL*, 2013.

[Misra *et al.*, 2015] Amita Misra, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. Using summarization to discover argument facets in online idealogical dialog. In *NAACL*, 2015.

[Olteanu *et al.*, 2017] Alexandra Olteanu, Onur Varol, and Emre Kiciman. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *CSCW*, 2017.

[Pavalanathan and Eisenstein, 2015] Umashanthi Pavalanathan and Jacob Eisenstein. Emoticons vs. emojis on twitter: A causal inference approach. *arXiv:1510.08480*, 2015.

[Pearl, 2009] Judea Pearl. *Causality*. 2009.

[Pennebaker *et al.*, 2007] James W Pennebaker, Roger J Booth, and Martha E Francis. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: LIWC.net*, 2007.

[Robins *et al.*, 1994] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, pages 846–866, 1994.

[Rosenthal and McKeown, 2015] Sara Rosenthal and Kathy McKeown. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *SIGDIAL*, 2015.

[Rubin, 2005] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, pages 322–331, 2005.

[Sridhar *et al.*, 2015] Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. Joint models of disagreement and stance in online debate. In *ACL*, 2015.

[Tan *et al.*, 2014] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. In *ACL*, 2014.

[Tan *et al.*, 2016] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *WWW*, 2016.

[Van der Laan and Rose, 2011] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[Veitch *et al.*, 2019] Victor Veitch, Yixin Wang, and David M Blei. Using embeddings to correct for unobserved confounding. *arXiv:1902.04114*, 2019.

[Walker *et al.*, 2012a] Marilyn Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martell, and Joseph King. That's your evidence?: Classifying stance in online political debate. *Decision Support Sciences*, 53(4), 2012.

[Walker *et al.*, 2012b] Marilyn Walker, Pranav Anand, Robert Abbott, and Jean E. Fox Tree. A corpus for research on deliberation and debate. In *LREC*, 2012.

[Walker *et al.*, 2012c] Marilyn Walker, Pranav Anand, Robert Abbott, and Richard Grant. Stance classification using dialogic properties of persuasion. In *NAACL*, 2012.

[Wei *et al.*, 2016] Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? ranking argumentative comments in online forum. In *ACL*, 2016.