Discovery of Treatments from Text Corpora

Christian Fong[*] Justin Grimmer[†]

March 3, 2016

Abstract

An extensive literature in computational social science examines how features of messages, advertisements, and other corpora affect individuals' decisions, but these analyses must specify the relevant features of the text before the experiment. Automated text analysis methods are able to discover features of text, but these methods cannot be used to obtain the estimates of causal effects—the quantity of interest for applied researchers. We introduce a new experimental design and statistical model to simultaneously discover treatments in a corpora and estimate causal effects for these discovered treatments. We prove the conditions to identify the treatment effects of texts and introduce the supervised Indian Buffet process to discover those treatments. Our method enables us to discover treatments in a training set using a collection of texts and individuals' responses to those texts, and then estimate the effects of these interventions in a test set of new texts and survey respondents. We apply the model to an experiment about candidate biographies, recovering intuitive features of voters' decisions and revealing a penalty for lawyers and a bonus for military service.

1 Introduction

Computational social scientists are often interested in inferring how blocks of text, such as messages from political candidates or advertising content, affect individuals' decisions (Ansolabehere and Iyengar, 1995; Mutz, 2011; Tomz and Weeks, 2013). To do so, they typically attempt to estimate the causal effect of the text: they model the outcome of interest, Y, as a function of the block of text presented to the respondent, t, and define the treatment effect of t relative to some other block

[*] Ph.D. Student, Graduate School of Business, Stanford University, cjfong@stanford.edu

[†] Associate Professor, Department of Political Science, Stanford University, JustinGrimmer.org, jgrimmer@stanford.edu.

of text t' as $Y(t) - Y(t')$ (Rubin, 1974; Holland, 1986). For example, in industrial contexts researchers design A/B tests to compare two potential texts for a use case. Academic researchers often design one text that has a feature of interest and another text that lacks that feature but is otherwise identical (for example, Albertson and Gadarian 2015). Both kinds of experiments assume researchers already know the features of text to vary and offer little help to researchers who would like to discover the features to vary.

Topic models and related methods can discover important features in corpora of text data, but they are constructed in a way that makes it difficult to use the discovered features to estimate causal effects (Blei, Ng and Jordan, 2003). Consider, for example, supervised latent Dirichlet allocation (sLDA) (Mcauliffe and Blei, 2007). It associates a topic-prevalence vector, $\boldsymbol{\theta}$, with each document where the estimated topics depend upon both the content of documents and a label associated with each document. If K topics are included in the model, then $\boldsymbol{\theta}$ is defined on the $K-1$-dimensional unit simplex, which makes treatment effects difficult to interpret. For example, suppose for two potential treatments $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$ that the observed response is $Y(\boldsymbol{\theta}) > Y(\boldsymbol{\theta}')$. Because topic-prevalence is defined on the unit-simplex, however, we are unsure if this is because for some topic j $\theta_j > \theta'_j$— an increased prevalence of topic j—or if it is because for some other topic k $\theta_k < \theta'_k$—a decreased prevalence of topic k. Fundamentally, this will be true for the output from all topic models because the zero-sum nature of the topic-prevalence vector implies that increasing the prevalence of any one topic necessarily decreases the prevalence of some other topic. The result is that it is difficult (or impossible) to interpret the effect of any one topic marginalizing over the other topics. Other applications of topic models to estimate causal effects treat text as the response, rather than the treatment (Roberts, Stewart and Airoldi, 2016). And still other methods require a difficult to interpret assumption of how text might affect individuals' responses (Beauchamp, 2011).

To facilitate the discovery of treatments and to address the limitation of existing unsupervised learning methods, we introduce a new experimental design, framework, and statistical model for discovering treatments within blocks of text and then reliably inferring the effects of those treatments. By doing so, we combine the utility of discovering important features in a topic model with the scientific value of causal treatment effects estimated in a potential outcomes framework. We present a new statistical model—the supervised Indian Buffet Process—to both discover treatments in a training set and infer the effects treatments in a test set (Ghahramani and Griffiths, 2005). We prove that randomly assigning blocks of text to respondents in an experiment is sufficient to identify the effects of latent treatments that comprise blocks of text.

Our framework provides the first of its kind approach to automatically discover treatment effects in text, building on literatures in both social science and machine learning (Blei, Ng and Jordan, 2003; Beauchamp, 2011; Mcauliffe and Blei, 2007; Roberts, Stewart and Airoldi, 2016). The use of the training and test set ensures that this discovery does not come at the expense of credibly inferring causal effects, insulating the research design from concerns about "p-hacking" and overfitting (Ioannidis, 2005; Humphreys, de la Sierra and van der Windt, 2013; Franco, Malhotra and Simonovits, 2014). Critically, we use a theoretical justification for our methodology: we select our particular approach because it enables us to estimate causal effect of interest. Rather than demonstrating that our method performs better at some predictive task, we prove that our method is able to estimate useful causal effects from the data.

We apply our framework to study how features of a political candidate's background affect voters' decisions. We use a collection of candidate biographies collected from Wikipedia to automatically discover treatments in the biographies and then infer their effects. This reveals a penalty for lawyers and career politicians and a bonus for military service and advanced degrees.

2 A Framework for Discovering Treatments from Text

Our goal is to discover a set of features—treatments—underlying texts and then estimate the effect of those treatments on some response from an individual. We first show that randomly assigning texts to respondents is sufficient to identify treatment effects. We then provide a statistical model for using both the text and responses to discover latent features in the text that affect the response. Finally, we show that we can use the mapping from text to features discovered on a training set to estimate the presence of features in a test set, which allows us to estimate treatment effects in the test set.

2.1 Randomizing Texts Identifies Underlying Treatment Effects

When estimating treatment effects, researchers often worry that the respondents who received one treatment systematically differ from those who received some other treatment. In a study of advertising, if all of the people who saw one advertisement were men and all of the people who saw a different advertisement were women, it would be impossible to tell whether differences in their responses were driven by the fact that they saw different advertisements or by their pre-existing differences. Randomized experiments are the gold standard for overcoming this problem (Gerber and Green, 2012). However, in text experiments, individuals are randomly assigned to blocks of text rather than to the latent features of the text that we analyze as the treatments. In this section, we show that randomly assigning blocks of text is sufficient to identify treatment effects.

To establish our result, we suppose we have a corpora of J texts, $\mathcal{X}$. We represent a specific text with $\boldsymbol{X}_j \in \mathcal{X}$, with $\boldsymbol{X}_j \in \Re^D$. We have a sample of N respondents from a population, with the response of individual i to text j given by the potential outcome $Y_i(\boldsymbol{X}_j)$. We suppose that for each document j there is a corresponding vector of K binary treatments $\boldsymbol{Z}_j \in \mathcal{Z}$ where $\mathcal{Z}$ contains all 2^K possible combinations of treatments, $\{0,1\}^K$. The function $g : \mathcal{X} \to \mathcal{Z}$ maps from the texts to the binary set of treatments: we will learn this function using the supervised Indian Buffet process introduced in the next section. Note that distinct elements of $\mathcal{X}$ may map to the same element of $\boldsymbol{Z}$.

To establish our identification result, we assume (Assumption 1) that $Y_i(\boldsymbol{X}_j) = Y_i(g(\boldsymbol{X}_j))$ for all $\boldsymbol{X}_j \in \mathcal{X}$ and all i, or that $\boldsymbol{Z}_j$ is sufficent to describe the effect of a document on individual i's response. Stated differently, we assume an individual would respond in the same way to two different texts if those texts have the same latent features. We further suppose (Assumption 2) that texts are randomly assigned to respondents according to probability measure h, ensuring that $Y_i(g(\boldsymbol{X}_j)) \perp\!\!\!\perp \boldsymbol{X}_j$ for all $\boldsymbol{X}_j \in \mathcal{X}$ and for all individuals i. This assumption ensures unobserved characteristics of individuals are not confounding inferences about the effects of texts. The random assignment of texts to individuals induces a distribution over a probability measure on treatment vectors $\boldsymbol{Z}$, $f(\boldsymbol{Z}) = \int_{\mathcal{X}} 1(\boldsymbol{Z} = g(\boldsymbol{X}))h(\boldsymbol{X})d\boldsymbol{X}$. Finally, we assume (Assumption 3) that $f(\boldsymbol{Z}) > 0$ for all $\boldsymbol{Z} \in \mathcal{Z}$.[1] This requires that every combination of treatment effects is possible from the documents in our corpus. In practice, when designing our study we want to ensure that the treatments are not *aliased* or perfectly correlated. If perfect correlation exists between factors, we are unable to disentangle the effect of individual factors.

In this paper we focus on estimating the Average Marginal Component Specific Effect for factor k (AMCE$_k$) (Hainmueller, Hopkins and Yamamoto, 2014). The AMCE$_k$ is useful for finding the effect of one feature, k, when k interacts with the other features in some potentially complicated way. It is defined as the difference in outcomes when the feature is present and when it is not present, averaged over the

[1] Note for this assumption to hold it is necessary, but not sufficient that g is a surjection from $\mathcal{X}$ onto $\mathcal{Z}$.

values of all of the other features. Formally,

$$\text{AMCE}_k = \int_{\boldsymbol{Z}_{-k}} \mathbb{E}\left[Y(Z_k=1,\boldsymbol{Z}_{-k}) - Y(Z_k=0,\boldsymbol{Z}_{-k})\right] m(\boldsymbol{Z}_{-k}) d\boldsymbol{Z}_{-k}$$

where $m(\boldsymbol{Z}_{-k})$ is some analyst-defined density on all elements but k of the treatment vector. For example, $m(\cdot)$ can be chosen as the density of Z_{-k} in the population to obtain the marginal component effect of k in the empirical population. The most commonly used $m(\cdot)$ in applied work is uniform across all Z_{-k}'s, and we follow this convention here.

We now prove that assumptions 1, 2, and 3 are sufficient to identify the AMCE$_k$ for all k.

Proposition 1. *Assumptions 1, 2, and 3 are sufficient to identify the AMCE$_k$ for arbitrary k.*

Proof. To obtain a useful form, we first marginalize over the documents,

$$\begin{aligned}\int_{\boldsymbol{Z}_{-k}}\int_{\mathcal{X}} \mathbb{E}\left[Y(Z_k=1,\boldsymbol{Z}_{-k})\right] f(\boldsymbol{Z}_{-k}|Z_k=1,\boldsymbol{X}) - \mathbb{E}\left[Y(Z_k=0,\boldsymbol{Z}_{-k})\right] f(\boldsymbol{Z}_{-k}|Z_k=0,\boldsymbol{X}) h(\boldsymbol{X}) d\boldsymbol{X} &= \\ \int_{\boldsymbol{Z}_{-k}} \mathbb{E}\left[Y(Z_k=1,\boldsymbol{Z}_{-k})\right] f(\boldsymbol{Z}_{-k}|Z_k=1) - \mathbb{E}\left[Y(Z_k=0,\boldsymbol{Z}_{-k})\right] f(\boldsymbol{Z}_{-k}|Z_k=0) d\boldsymbol{Z}_{-k} & \end{aligned} \tag{2.1}$$

If $f(\boldsymbol{Z}_{-k}|Z_k=0) = f(\boldsymbol{Z}_{-k}|Z_k=1) = m(\boldsymbol{Z}_{-k})$ then Equation 2.1 is the AMCE$_k$. Otherwise consider $m(\boldsymbol{Z}) > 0$ for all $\boldsymbol{Z} \in \mathcal{Z}$. Then because $f(\boldsymbol{Z}) > 0$ then $f(\boldsymbol{Z}_{-k}|Z_k=0) > 0$ and $f(\boldsymbol{Z}_{-k}|Z_k=1) > 0$. Thus, there exists conditional densities $h(\boldsymbol{Z}|Z_k=1)$ and $h(\boldsymbol{Z}|Z_k=0)$ such that $\frac{f(\boldsymbol{Z}_{-k}|Z_k=1)}{h(\boldsymbol{Z}_{-k}|Z_k=1)} = \frac{f(\boldsymbol{Z}_{-k}|Z_k=0)}{h(Z_{-k}|Z_k=0)} = m(\boldsymbol{Z}_{-k})$ □

2.2 A Statistical Model for Identifying Features

The preceding section shows that if we are able to discover features in the data, we can estimate their AMCEs by randomly assigning blocks of texts to respondents. We now present a statistical model for discoering those features. As we argued in the introduction, it is difficult to use the topics obtained from topic models like sLDA because the topic vector exists on the simplex. When we compare the outcomes associated with two different topic vectors, we do not know whether the change in the response is caused by increasing the degree to which the document about one topic or decreasing the degree to which it is about another, because the former mathematically entails the latter. Other models, such as LASSO regression, would necessarily suppose that the presence and absence of words are the treatments (Hastie, Tibshirani and Friedman, 2001; Beauchamp, 2011). This is problematic substantively, because it is hard to know exactly what the presence or absence of a single word implies as a treatment in text.

We therefore develop the supervised Indian Buffet Process (sIBP) to discover features in the document. For our purposes, the sIBP has two essential properties. First, it produces a binary topic vector, avoiding the complications of treatments assigned on the simplex. Second, unlike the Indian Buffet Process upon which it builds (Ghahramani and Griffiths, 2005), it incorporates information about the outcome associated with various texts, and therefore discovers features that explain both the text and the response.

Figure 2.2 describes the posterior distribution for the sIBP and a summary of the posterior is given in Equation 2.2. We describe the model in three steps: the treatment assignment process, document creation, and response. The result is a model that creates a link between document content and response through a vector of treatment assignments.

Treatment Assignment We assume that $\boldsymbol{\pi}$ is a K-vector (where we take the limit as $K \to \infty$), where π_k describes the population proportion of documents that exhibit latent feature k. For document j and topic k we suppose that $z_{j,k} \sim \text{Bernoulli}(\pi_k)$, which importantly implies that the occurrence of treatments are not zero sum. We collect the treatment vector for document j into $\boldsymbol{Z}_j$ and collect all the treatment vectors into $\mathbf{Z}$ an $N_{\text{texts}} \times K$ binary matrix, where N_{texts} refers to number of unique documents. Throughout we will assume that $N_{\text{texts}} = N$ or that the number of documents and responses are equal and index the documents with i.

Document Creation We suppose that the documents are created as a combination of latent factors. For topic k we suppose that $\boldsymbol{A}_k$ is a $D-$dimensional vector that maps latent features onto observed text. We collect the vectors into $\boldsymbol{A}$, a $K \times D$ matrix, and suppose that $\boldsymbol{X}_i \sim \text{MVN}(\boldsymbol{Z}_i \boldsymbol{A}, \sigma_n^2 I_D)$.

Response to Treatment Vector We assume that a $K-$vector of parameters $\boldsymbol{\beta}$ describes the relationship between the treatment vector and response. Specifically, we use a standard parameterization and suppose that $\tau \sim \text{Gamma}(a, b)$, $\boldsymbol{\beta} \sim \text{MVN}(0, \tau^{-1})$ and that $Y_i \sim \text{Normal}(\boldsymbol{Z}_i \boldsymbol{\beta}, \tau^{-1})$.

$$\begin{aligned}
\pi_k &\sim \text{Beta}\left(\frac{\alpha}{K}, 1\right) \\
z_{i,k} &\sim \text{Bernoulli}(\pi_k) \\
\mathbf{X}_i | \mathbf{Z}_i, \mathbf{A} &\sim \text{MVN}(\mathbf{Z}_i \mathbf{A}, \sigma_X^2 I_D) \\
\mathbf{A}_k &\sim \text{MVN}(\mathbf{0}, \sigma_A^2 I_D) \\
\mathbf{Y}_i | \mathbf{Z}_i, \boldsymbol{\beta} &\sim \text{Normal}(\mathbf{Z}_i \boldsymbol{\beta}, \tau^{-1}) \\
\tau &\sim \text{Gamma}(a, b) \\
\boldsymbol{\beta} | \tau &\sim \text{MVN}(\mathbf{0}, \tau^{-1} I_K) \qquad (2.2)
\end{aligned}$$

2.2.1 Inference for the Supervised Indian Buffet Process

We approximate the posterior distribution with a variational approximation, building on the algorithm introduced in (Doshi-Velez et al., 2009). We approximate the non-parametric posterior setting K to be large and use a factorized approximation,

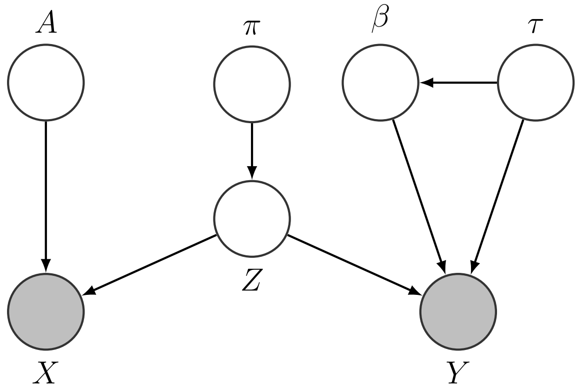

Figure 1: Graphical Model for the Supervised Indian Buffet Process

assuming that

$$p(\boldsymbol{\pi}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\beta}, \tau | \boldsymbol{X}, \boldsymbol{Y}, \alpha, \sigma_A^2, \sigma_X^2, a, b) \;\; = \;\; q(\boldsymbol{\pi})q(\boldsymbol{A})q(\boldsymbol{Z})q(\boldsymbol{\beta}, \tau)$$

A standard derivation that builds on Doshi-Velez et al. (2009) leads to the following distributional forms and update steps:

- $q(\pi_K) = \text{Beta}(\pi_k|\boldsymbol{\lambda}_k)$. The update values are $\lambda_{k,1} = \frac{\alpha}{K} + \sum_{i=1}^{N} \nu_{i,k}$ and $\lambda_{k,2} = 1 + \sum_{i=1}^{N}(1 - \nu_{i,k})$.

- $q(\boldsymbol{A}_k) = \text{Multivariate Normal}(\boldsymbol{A}_k|\bar{\boldsymbol{\phi}}_k, \boldsymbol{\Phi}_k)$. The updated parameter values are,

$$\begin{aligned} \bar{\phi}_k &= \left[\frac{1}{\sigma_X^2} \sum_{i=1}^{N} \nu_{i,k} \left(\mathbf{X}_i - \left(\sum_{l:l \neq k} \nu_{n,l} \bar{\phi}_l \right) \right) \right] \left(\frac{1}{\sigma_A^2} + \frac{\sum_{i=1}^{N} \nu_{i,k}}{\sigma_X^2} \right)^{-1} \\ \boldsymbol{\Phi}_k &= \left(\frac{1}{\sigma_A^2} + \frac{\sum_{i=1}^{N} \nu_{i,k}}{\sigma_X^2} \right)^{-1} I \end{aligned}$$

- $q(\boldsymbol{\beta}, \boldsymbol{\tau}) = \text{Multivariate Normal}(\boldsymbol{\beta}, \boldsymbol{m}, \boldsymbol{S}) \times \text{Gamma}(\tau|c, d)$. The updated parameter values are,

$$
\begin{aligned}
\boldsymbol{m} &= \mathbf{S}\mathbb{E}[\mathbf{Z}^T]\mathbf{Y} \\
\boldsymbol{S} &= (\mathbb{E}[\mathbf{Z}^T\mathbf{Z}] + I_K)^{-1} \\
c &= a + \frac{N}{2} \\
d &= b + \frac{\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbb{E}[\mathbf{Z}]\mathbf{S}\mathbb{E}[\mathbf{Z}^T]\mathbf{Y}}{2}
\end{aligned}
$$

Where typical element of $\mathbb{E}[\boldsymbol{Z}^T]_{j,k} = \nu_{j,k}$ and typical on-diagonal element of $\mathbb{E}[\boldsymbol{Z}^T\boldsymbol{Z}]_{k,k} = \sum_{i=1}^N \nu_{i,k}$ and off-diagonal element is $\mathbb{E}[\boldsymbol{Z}^T\boldsymbol{Z}]_{j,k} = \sum_{i=1}^N \nu_{i,j}\nu_{i,k}$.

- $q(z_{i,k}) = \text{Bernoulli}(z_{i,k}|\nu_{i,k})$. The updated parameter values are

$$
\begin{aligned}
\upsilon_{i,k} &= \psi(\lambda_{k,1}) - \psi(\lambda_{k,2}) - \frac{1}{2\sigma_X^2}\left[-2\bar{\phi}_k\mathbf{X}_i^T + (tr(\mathbf{\Phi}_k) + \bar{\phi}_k\bar{\phi}_k^T) + 2\bar{\phi}_k\left(\sum_{l:l\neq k}\nu_{i,l}\bar{\phi}_l^T\right)\right] \\
&\quad -\frac{c}{2d}\left(-2m_kY_i + \left(\frac{dS_{k,k}}{c-1} + m_k^Tm_k\right) + 2m_k\left(\sum_{l:l\neq k}\nu_{i,l}m_l\right)\right) \\
\nu_{i,k} &= \frac{1}{1+\exp(-\upsilon_{i,k})} \qquad (2.3)
\end{aligned}
$$

where $\psi(\cdot)$ is the digamma function. We repeat the algorithm until the change in the parameter vector drops below a threshold.

To select the final model using the training set data, we run the algorithm several times using a procedure we developed to search over values of α and σ_X.[2] We then run the model several times for each combination of values for α and σ_X to evaluate the output at several different local modes. To create a candidate set of models, we use a quantitative measure that balances coherence and exclusivity (Mimno et al., 2011; Roberts et al., 2014). We identify the top ten words for intervention k as the ten words with the largest value in $\boldsymbol{A}_k$, $\boldsymbol{t}_k$ and define $N_k = \sum_{i=1}^N \nu_{i,k}$. We then

[2] We assign σ_A, a, and b values which lead to diffuse priors.

obtain measure CE for a particular model,

$$\text{CE} \;=\; \sum_{k=1}^{K} N_k \sum_{l,c \in t_k} \text{cov}(\boldsymbol{\nu}_k \boldsymbol{X}_l, \boldsymbol{\nu}_k \boldsymbol{X}_c) - \sum_{k=1}^{K} (N - N_k) \sum_{l,c \in t_k} \text{cov}((\mathbf{1} - \boldsymbol{\nu}_k)\boldsymbol{X}_l, (\mathbf{1} - \boldsymbol{\nu}_k)\boldsymbol{X}_c)$$

We make a final model selection based on the model that provides the most substantively clear treatments.

2.3 Inferring Treatments and Estimating Effects in Test Set

To discover the treatment effects, we first suppose that we have randomly assigned a set of respondents a text based treatment $\boldsymbol{X}_i$ according to some probability measure $h(\cdot)$ and that we have observed their response Y_i. We collect the assigned texts into $\boldsymbol{X}$ and the responses into $\boldsymbol{Y}$. As we describe below, we will often assign each respondent their own distinctive message, with the probability of receiving any one message at $\frac{1}{N}$ for all respondents and messages. We use the sIBP model trained our training set documents and responses to infer the effect of those treatments among the test set documents. To make this inference, we first use the model from the training set to infer the posterior distribution on treatment vectors for documents in the test set.

We approximate the posterior distribution for the treatment vectors using the variational approximation from the training set parameters $(\widehat{\boldsymbol{\lambda}}, \widehat{\bar{\boldsymbol{\phi}}}, \widehat{\boldsymbol{\Phi}}, \widehat{\boldsymbol{m}}, \widehat{\boldsymbol{S}}, \widehat{c}, \widehat{d}, \widehat{\sigma_X^2}, \widehat{\sigma_A^2})$ and a modified update step on $q(z_{i,k}^{\text{test}})$. In this modified update step, we remove the component of the update that incorporates information about the outcome. Specifically for individual i in the test set for category k we have the following update step

$$\begin{aligned} v_{i,k}^{\text{test}} &= \psi(\widehat{\lambda_{k,1}}) - \psi(\widehat{\lambda_{k,2}}) - \frac{1}{2(\widehat{\sigma_X^2})}\left[-2\widehat{\bar{\phi}_k}(\mathbf{X}_i^T) + (tr(\widehat{\boldsymbol{\Phi}_k}) + \widehat{\bar{\phi}_k}(\widehat{\bar{\phi}_k})^T) + 2\widehat{\bar{\phi}_k}\left(\sum_{l:l\neq k} \nu_{i,l}\left(\widehat{\bar{\phi}_l}^T\right)\right)\right] \\ \nu_{i,k}^{\text{test}} &= \frac{1}{1+\exp(-v_{i,k}^{\text{test}})}. \end{aligned} \tag{2.4}$$

For each text in the test set we repeat this update several times until $\boldsymbol{\nu}^{\text{test}}$ has converged. Note that Equation 2.4 differs from Equation 2.3 in that the component of the model that generates the response for the test set respondents is excluded.

With the approximating distribution $q(\boldsymbol{Z}^{\text{test}})$ we then measure the effect of the treatments in the test set. Using the treatments, the most straightforward model to estimate assumes that there are no interactions between each of the components. Under the no interactions assumption, we estimate the effects of the treatments and infer confidence intervals using the following bootstrapping procedure that incorporates uncertainty both from estimation of treatments and uncertainty about the effects of those treatments:

1) For each respondent i and component k we draw $\tilde{z}_{i,k} \sim \text{Bernoulli}(\nu_{i,k}^{\text{test}})$, resulting in vector $\tilde{\boldsymbol{Z}}$

2) Given the vector $\tilde{Z}$, we sample $(\boldsymbol{Y}^{\text{test}}, \tilde{\boldsymbol{Z}})$ with replacement and for each sample estimate the regression $\boldsymbol{Y}^{\text{test}} = \boldsymbol{\beta}^{\text{test}}\tilde{\boldsymbol{Z}} + \boldsymbol{\epsilon}$.

We repeat the bootstrap steps 1000 times, keeping $\boldsymbol{\beta}^{\text{test}}$ for each iteration. The result of the procedure is a point estimate of the effects and confidence interval of the treatments under no interactions.

3 Application: Voter Evaluations of an Ideal Candidate

We demonstrate our method in an experiment to assess how features of a candidate's background affect respondents evaluations of the candidates. There is a rich literature in political science about the ideal attributes of political candidates (Canon, 1990; Popkin, 1994; Carnes, 2012; Campbell and Cowley, 2014). We build on this literature and use a collection of candidate biographies to discover features of candidates' backgrounds that voters find appealing. To uncover the features of candidate biographies that voters are responsive to we acquired a collection of 1,246 Congressional candidate biographies from Wikipedia. We then anonymize the biographies—replacing names and removing other identifiable information—to ensure that the only information available to the respondent was explicitly present in the text.

In Section 2.1 we show that a necessary condition for this experiment to uncover latent treatments is that each vector of treatments has non-zero probability of occuring. This is equivalent to assuming that none of the treatments are *aliased*, or perfectly correlated (Hainmueller, Hopkins and Yamamoto, 2014). Aliasing would be more likely if there are only a few distinct texts that are provided to participants in our experiment. Therefore, we assign each respondent in each evaluation round a distinct candidate biography. To bolster our statistical power, we ask our respondents to evaluate up to four distinct candidate biographies, resulting in each respondent evaluating 2.8 biographies on average.[3] After presenting the respondents with a candidate's biography, we ask each respondent to rate the candidate using a *feeling thermometer*: a well-established social science scale that goes from 0 when a respodent is "cold" to a candidate to 100 when a respondent is "warm" to the candidate.

We recruited a sample of 1,886 participants using Survey Sampling International (SSI), an online survey platform. Our sample is census matched to reflect US demographics on sex, age, race, and education. Using the sample we obtain 5,303 total observations. We assign 2,651 responses to the training set and 2,652 to the test set. We then apply the sIBP process to the training data. To apply the model, we standardize the feeling thermometer to have mean zero and standard deviation 1. We set K to a relatively low value ($K = 10$) reflecting a quantitative and qualitative

[3] The multiple evaluations of candidate biographies is problematic if there is spillover across rounds of our experiment. We have little reason to believe observing one candidate biography would systematically affect the response in subsequent rounds.

Table 1: Top Words for 10 Treatments sIBP Discovered

Treatment 1	**Treatment 2**	**Treatment 3**	**Treatment 4**	**Treatment 5**
appointed	fraternity	director	received	elected
school_graduated	distinguished	university	washington_university	house
governor	war_ii	received	years	democratic
worked	chapter	president	death	seat
older	air_force	master_arts	company	republican
law_firm	phi	phd	training	served
elected	reserve	policy	military	committee
grandfather	delta	public	including	appointed
office	air	master	george_washington	defeated
legal	states_air	affairs	earned_bachelors	office
Treatment 6	**Treatment 7**	**Treatment 8**	**Treatment 9**	**Treatment 10**
united_states	republican	star	law	war
military	democratic	bronze	school_law	enlisted
combat	elected	germany	law_school	united_states
rank	appointed	master_arts	juris_doctor	assigned
marine_corps	member	awarded	student	army
medal	incumbent	played	earned_juris	air
distinguished	political	yale	earned_law	states_army
air_force	father	football	law_firm	year
states_air	served	maternal	university_school	service
air	state	division	body_president	officer

search over K. We then select the final model varying the parameters and evaluating the CE score.

Table 1 provides the top words for each of the ten treatments the sIBP discovered in the training set. The treatments cover salient features of Congressional biographies from the time period that we analyze. For example, treatments 6 and 10 capture a candidate's military experience. Treatment 5 and 7 are about previous political experience and Treatment 3 and 9 refer to a candidate's education experience. Obviously, there are many features of a candidate's background missing here, but the treatments discovered provide a useful set of dimensions to assess how voters respond to a candidate's background.

After training the model on the training set, we apply it to the test set to infer the treatments in the biographies. We assume there are no interactions between

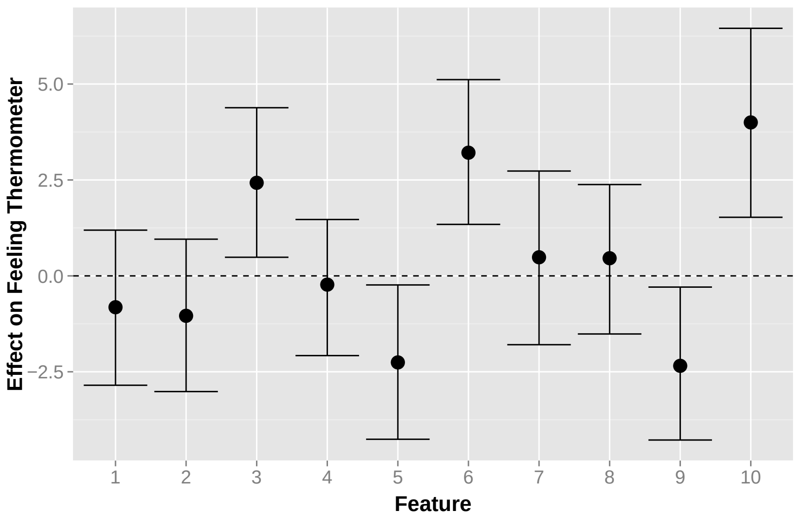

Figure 2: 95% Confidence Intervals for Effects of Discovered Treatments: The mean value of the feeling thermometer is 62.3

the discovered treatments in order to estimate their effects.[4] Figure 2 shows the point estimate and 95-percent confidence intervals, which take into account uncertainty in inferring the treatments from the texts and the relationship between those treatments and the response.

The treatment effects reveal intuitive, though interesting, features of candidate biographies that affect respondent's evluations. For example, Figure 2 reveals a distaste for political and legal experience—even though a large share of Congressional candidates have previous political experience and a law degree. Treatment 5, which describes a candidate's previous political experience, causes an 2.26 point reduction in feeling thermometer evaluation (95 percent confidence interval, [-4.26,-0.24]). Likewise, Treatment 9 shows that respondents dislike lawyers, with the presence of legal experience causing a 2.34 point reduction in feeling thermometer (95-percent confidence interval, [-4.28,-0.29]). The aversion to lawyers is not, however, an aversion to education. Treatment 3, a treatment that describes advanced degrees, causes a 2.43 point increase in feeling thermometer evaluations (95-percent confidence in-

[4] This assumption is not necessary for the framework we propose here. Interaction effects could be modeled.

terval, [0.49,4.38]).

In contrast, Figure 2 shows that there is a consistent bonus for military experience. This is consistent with intuition from political observers that the public supports veterans. For example, treatment 6, which describes a candidate's military record, causes a 3.21 point increase in feeling thermometer rating (95-percent confidence interval, [1.34,5.12]) and treatment 10 causes a 4.00 point increase (95-percent confidence interval, [1.53,6.45]).

4 Conclusion

We have presented a methodology for discovering treatments in text and then inferring the effect of those treatments on respondents' decisions. We prove that randomizing texts is sufficient to identify the underlying treatments and introduce the supervised Indian Buffet process for discovering the effects. The use of a training and test set ensures that our method provides accurate confidence intervals and avoids the problems of overfitting or "p-hacking" in experiments. In an application to candidate biographies, we discover a penalty for political and legal experience and a bonus for military service and non-legal advanced degrees.

Our methodology has a wide variety of applications. This includes numerous alternative experimental designs, providing a methodology that computational social scientists could use widely to discover and then confirm the effects of messages in numerous domains—including images and other high dimensional data. The methodology is also useful for observational data—for studying the effects of complicated treatments, such as how a legislator's roll call voting record affects their electoral support.

References

Albertson, Bethany and Shana Kushner Gadarian. 2015. *Anxious Politics: Democratic Citizenship in a Threatening World.* Cambridge University Press.

Ansolabehere, Stephen and Shanto Iyengar. 1995. *Going Negative: How Political Advertisements Shrink and Polarize The Electorate.* Simon & Schuster, Inc.

Beauchamp, Nick. 2011. "A Bottom-up Approach to Linguistic Persuasion in Advertising." *The Political Methodologist* .

Blei, David, Andrew Ng and Michael Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning and Research* 3:993–1022.

Campbell, Rosie and Philip Cowley. 2014. "What Voters Want: Reactions to Candidate Characteristics in a Survey Experiment." *Political Studies* 62(4):745–765.

Canon, David T. 1990. *Actors, Athletes, and Astronauts: Political Amateurs in the United States Congress.* University of Chicago Press.

Carnes, Nicholas. 2012. "Does the Numerical Underrepresentation of the Working Class in Congress Matter?" *Legislative Studies Quarterly* 37(1):5–34.

Doshi-Velez, Finale, Kurt T. Miller, Jurgen Van Gael and Yee Whye Teh. 2009. "Variational Inference for the Indian Buffet Process." Technical Report, University of Cambridge.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203):1502–1505.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiment: Design, Analysis, and Interpretation.* W.W. Norton & Company.

Ghahramani, Zoubin and Thomas L Griffiths. 2005. Infinite Latent Feature Models and the Indian Buffet Process. In *Advances in neural information processing systems.* pp. 475–482.

Hainmueller, Jens, Daniel Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multi-Dimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2001. *The Elements of Statistical Learning.* Springer.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21(1):1–20.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings are False." *PLoS Medicine* 2(8):696–701.

Mcauliffe, Jon D. and David M. Blei. 2007. Supervised Topic Models. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*.

Mimno, David, Hanna M Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp. 262–272.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton University Press.

Popkin, Samuel L. 1994. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Chris Lucas, Jetson Leder-Luis, Bethany Albertson, Shana Gadarian and David Rand. 2014. "Topic Models for Open Ended Survey Responses with Applications to Experiments." *American Journal of Political Science* .

Roberts, Margaret E., Brandon M. Stewart and Edo M. Airoldi. 2016. "A model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* . Forthcoming.

Rubin, Don. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701.

Tomz, Michael and Jessica Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107(4):849–865.