Estimating the causal effects of treatments using Yelp reviews

Luisa Quispe Ortiz[1] **and Varun D N**[1]

[1]Center for Data Science, New York University
[*]lqo202@nyu.edu, vdn207@nyu.edu

ABSTRACT

Exploring the latent factors influencing the decision of an individual is an interesting problem which enables us to reason their decision. Social scientists examine the effect of contents on individuals' decisions by conducting experiments involving hand-engineered features known to them beforehand. Automated text analysis methods allow us to identify features from a given textual corpora. But, these techniques do not allow for exploring the causal effects of these features. Fong et al.[1] propose a methodology involving a graphical model to simultaneously discover treatments (topics) from text along with the causal effects of them on the individuals' response. In this work, we use their methodology on Yelp reviews to estimate the causal effects of user expectations on the ratings they provide. We use the model to understand the causal effects on the reviews given for Mexican restaurants in Nevada and reviews from popular cities in UK, Germany, Canada and US.

1 Introduction

Social scientists are interested in understanding how textual content influences individuals' decisions. Lot of work has been done to discover the factors that might have influenced the behavior of people based on their responses to contents. But, such techniques do not take the actual contents presented to the individuals into consideration. To estimate the causal effect of contents, researchers had initially proposed a methodology where they model the outcome of interest as a function of the different contents presented and the difference in the outcomes would represent the effect of contents[2][3]. A simple example of this methodology is the A/B testing framework. But, these methodologies involve engineering or knowing the features of textual content and offer less flexibility towards varying or dynamically inferring the features.

One prevalent technique to discover topics from textual corpora is Latent Dirichlet Allocation (LDA)[4]. Though the model allows us to learn topics from a text corpus, they are not designed to estimate the causal effects of the discovered topics. Supervised LDA[5] is another technique which takes the label of a document and the contents of it into consideration to learn the topic distribution for that document. The values of the topic distribution are relative i.e. the prevalence of any one topic necessarily decreases the prevalence of another topic. So, it is not clear how we can estimate the causal effect of any one topic marginalizing over the other topics.

To facilitate the discovery of treatments and to address the limitation of existing unsupervised learning methods, Fong et. al[1] introduced a new experimental design, framework and statistical model for discovering treatments (topics) within blocks of text and then reliably inferring the effects of those treatments. They present a new statistical model - the Supervised Indian Buffet Process - to both discover the treatments from a training set and infer the effects of treatments on a test set. They applied their framework to study the impact of a political candidate's background on voters' decisions. They discover that voters' are more impressed by candidates with military service and advanced degrees and are less impressed by candidates who were lawyers and career politicians.

In this work, we achieve two objectives. First, we replicate the analysis carried out by the authors and introspect into the results we obtain. Second, we use the methodology proposed by Fong et al. to understand the causal effects of treatments discovered from Yelp on the ratings given for a review. This will help us understand the factors that influence the rating given by a user and what people expect at a restaurant to enjoy their food. We found that customers give a higher rating when they were satisfied with the food taste and when the restaurant provided them with varieties of toppings and side-dishes. The customers give a bad rating when they get a bad service or were made to wait for a long time.

The report is structured in the following way: Section 2 explains the framework proposed by Fong et al. Section 3 has details about the replication process we conducted on the same dataset used by Fong et al. Section 4 has details about our experiment with Yelp reviews and contains a discussion of the results on them. We finally conclude and describe future work in Section 5.

2 Framework

2.1 Experimental Protocol

The experimental procedure followed by Fong et al. is as follows:[1]

1. Randomly assign text X_j
2. Obtain response Y_i for each respondent
3. Divide the texts and responses into training and test set
4. In training set:
    - Use supervised Indian Buffet Process (sIBP) applied to documents and responses to infer latent treatments in texts
    - Model selection via quantitative fit and qualitative assessment
5. In test set:
    - Use sIBP found in training set to infer latent treatments on test set documents
    - Estimate the effect of treatments with regression, with a bootstrap procedure to estimate uncertainty

2.2 Framework Explanation

2.2.1 Supervised Indian Buffet Process(sIBP)

The problem in hand is to find causal effects on topics (treatments). Due to this aim, using supervised LDA was not a good idea, because for a model with k topics the topic vector probabilities lay in a (k-1) simplex. This makes the analysis difficult because a change in the response can be due to increasing the distribution of one topic or decreasing the distribution of another topic [[1]].

Therefore, the authors decided to use sIBP because it avoids the simplex and it's possible to incorporate information associated with many texts allowing us to find features that explain the response and the text. It is worth pointing out that the model Fong et al used differs from the initial sIBP proposed in Quadrianto et al[6], leading to different ways to gather data and infer the parameters. As is indicated in their paper, the main difference is that Quadrianto's outcome is a preference relation tuple, while theirs is a real-valued scalar.

Supervised Indian Buffet can be explained in three main steps. Consider we want to use K topics and we have the text $\mathbf{X}$ that is data matrix of D variables and N observations, and correspondingly $\mathbf{Y}$ a N- dimensional vector.

- **Treatment Assignment** Consider π is a K-vector where π_k describes the population proportion of documents that contain latent feature k. Also, π is generated by the stick-breaking construction. This would be the topic vector sampled from a Dirichlet distribution in a classic LDA. The stick-breaking constructions uses η as a parameter, specifically, we suppose that $\eta_k \sim \text{Beta}(\alpha, 1)$ for all K. By definition also we have that $\pi_1 = \eta_1$ and for each remaining topic, we assume that $\pi_k = \prod_{z=1}^{k} \eta_z$. For document j and topic k, in LDA we will have an indicator variable, in this case there is that one too: $z_{j,k} \sim \text{Bernoulli}(\pi_k)$. Now, collecting all the treatment for document j, a K-dimensional vector Z_j is formed, furthermore if all document vectors are concatenated we form the matrix $\mathbf{Z}$ of dimension $N \times K$ binary matrix.
- **Document Creation** It's assumed that the reviews are created as a combination of latent factors. For topic k we suppose that A_k is a $D-$dimensional vector that maps latent features onto observed text. We collect the vectors into A, a $K \times D$ matrix, and suppose that $X_i \sim \text{MVN}(Z_i A, \sigma_n^2 I_D)$, where $X_{i,d}$ is the standardized number of times word d appears in review i. While it is common to model texts as draws from multinomial distributions, the multivariate normal distribution is useful for our purposes for two reasons. First, the data is normalized by transforming each column $X_{\cdot,d}$ to be mean 0 and variance 1, ensuring that the multivariate normal distribution captures the overall contours of the data. Note that this implies that $X_{i,d}$ can be negative. Second, Fong et al. show that assuming a multivariate normal for document generation results in parameters that capture the distinctive rate words are used for each latent feature

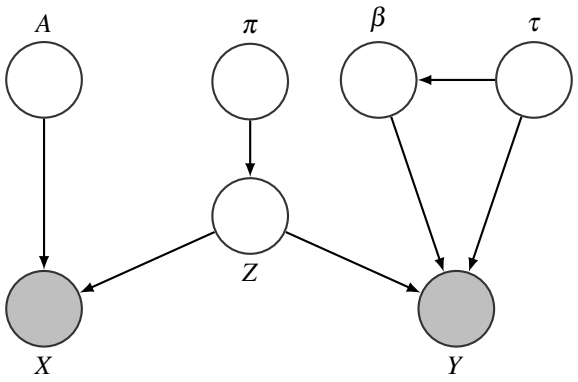

Figure 1. Graphical Model for the Supervised Indian Buffet Process [Source: Fong et al.[1]]

- **Response to Treatment Vector** Response is influenced by the treatment vector by Z and by another prior. The first one indicates the relationship with the $\mathbf{X}$ matrix, the other one will indicate the relationship with the treatment vector directly. Assume β is a $K-$vector of parameters that describes the relationship between the treatment vector and response with distribution $\beta \sim \text{MVN}(\mathbf{0}, \tau^{-1} I_K)$. The parameter of the MVN has prior over the variance matrix set to $\tau \sim \text{Gamma}(a,b)$. The relation to the response then is indicated by $Y_i \sim \text{Normal}(Z_i\beta, \tau^{-1})$.

$$
\begin{aligned}
\pi_k &\sim \text{Stick-Breaking}\,(\alpha) \\
z_{i,k} &\sim \text{Bernoulli}(\pi_k) \\
\mathbf{X}_i|\mathbf{Z}_i, \mathbf{A} &\sim \text{MVN}(\mathbf{Z}_i\mathbf{A}, \sigma_X^2 I_D) \\
\mathbf{A}_k &\sim \text{MVN}(\mathbf{0}, \sigma_A^2 I_D) \\
\mathbf{Y}_i|\mathbf{Z}_i, \beta &\sim \text{Normal}(\mathbf{Z}_i\beta, \tau^{-1}) \\
\tau &\sim \text{Gamma}(a,b) \\
\beta|\tau &\sim \text{MVN}(\mathbf{0}, \tau^{-1} I_K)
\end{aligned}
\tag{1}
$$

2.2.2 Variational Inference for sIBP

Doshi-Velez et al.[7] propose variational inference for the Indian Buffet Process. Since the prior on Z is non-parametric, it allows for the number of features to expand. With N documents and even limiting the number of features to K, the total number of assignments possible are $O(2^{NK})$ which is considerably large. This makes sampling based techniques really slow. Also, hard assignment gives the samplers less flexibility in moving between optima and might get stuck at different local optima. Variational inference approach allows for soft assignment and hence provides more flexibility than the sampling technique.

In the original paper, they propose two mean field approximations and in both the techniques, they limit the number of features that can be discovered. The first approach is called the finite variational approach where they minimize the KL-Divergence between the variational distribution and a finite approximation p_K to the IBP. The second approach is called the infinite variational approach where the KL-Divergence is minimized between the variational distribution and the true IBP posterior.

Building upon the infinite variational approach proposed above, Fong et al.[1] derive the following distributional forms an update steps:

- $q(\pi_K) = \text{Beta}(\pi_k|\lambda_k)$. The update values are $\lambda_{k,1} = \frac{\alpha}{K} + \sum_{i=1}^{N} \nu_{i,k}$ and $\lambda_{k,2} = 1 + \sum_{i=1}^{N}(1-\nu_{i,k})$.
- $q(A_k) = \text{Multivariate Normal}(A_k|\bar{\phi}_k, \Phi_k)$. The updated parameter values are,

$$\bar{\phi}_k = \left[\frac{1}{\sigma_X^2}\sum_{i=1}^{N} \nu_{i,k}\left(\mathbf{X}_i - \left(\sum_{l:l\neq k} \nu_{i,l}\bar{\phi}_l\right)\right)\right]\Phi_k$$

$$\Phi_k = \left(\frac{1}{\sigma_A^2} + \frac{\sum_{i=1}^N \nu_{i,k}}{\sigma_X^2}\right)^{-1} I$$

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

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

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

- $q(z_{i,k}) = \text{Bernoulli}(z_{i,k}|\nu_{i,k})$. The updated parameter values are
  $\nu_{i,k} = \psi(\lambda_{k,1}) - \psi(\lambda_{k,2}) - \frac{1}{2\sigma_X^2}[-2\bar{\phi}_k\mathbf{X}_i^T + (\text{tr}(\Phi_k) + \bar{\phi}_k\bar{\phi}_k^T) + 2\bar{\phi}_k\left(\sum_{l:l\neq k}\nu_{i,l}\bar{\phi}_l^T\right)]$
  $-\frac{c}{2d}(-2m_kY_i + \left(\frac{dS_{k,k}}{c-1} + m_k^Tm_k\right) +$
  $2m_k\left(\sum_{l:l\neq k}\nu_{i,l}m_l\right))$
  $\nu_{i,k} = \frac{1}{1+\exp(-\nu_{i,k})}$
  where $\psi(\cdot)$ is the digamma function.

This algorithm is repeated until the change in the parameter vector drops below a threshold.

To choose a final model, 10 iterations are performed for each combination of $\alpha \in \{3,5,8\}$ and $\sigma_X \in \{0.2,0.4,0.6\}$. For the parameters learned during each iteration of a configuration, a quantitative measure called the CE Score is used.

CE Score is computed as follows: Let $\mathscr{I}_k$ be the set of documents for which $\nu_{i,k} \geq 0.5$, and let $\mathscr{I}_k^C$ be the complement of this set. We identify the top ten words for intervention k as the ten words with the largest value in A_k, t_k and define

$$N_k = \sum_{i=1}^N I\{\nu_{i,k} \geq 0.5\}$$

. We then obtain measure CE for a particular model

$$\text{CE} = \sum_{k=1}^K N_k \sum_{l,c\in t_k} \text{cov}(X_{\mathscr{I}_k,l}, X_{\mathscr{I}_k,c}) - \sum_{k=1}^K (N-N_k) \sum_{l,c\in t_k} \text{cov}(X_{\mathscr{I}_k^C,l}, X_{\mathscr{I}_k^C,c})$$

where here $X_{\mathscr{I}_k,l}$ refers to the l^{th} column and $\mathscr{I}_k$th rows of X

2.3 Fong et al.'s Experiment

The work aimed to find how features of a candidate's background affect a voter's evaluation of the candidate. A collection of 1,246 candidate biographies was prepared from Wikipedia and the identifiers were anonymized in order to ensure the respondent had only candidate's background details available. Then, using Survey Sampling International (SSI), they performed a survey where the respondent had to read candidate biographies and rate (our Y target variable) the candidate using a feeling thermometer whose value is between 0-100 (0 being least likely to vote). They had 1,886 participants and each responded to 2.8 biographies on average leading to 5,303 observations.

To recognize the main unigrams and bigrams, the procedure first constructed a dictionary of words among all the documents (not considering the biographies identifier's). This resulted in about 32,000 words and selected the top 2% using a fixed threshold of 0.015 (average word per document count). With the common words, bigrams were constructed only if their words were in the common dictionary. To filter the most common bigrams a threshold of 0.015 in the count was imposed too, resulting in 164. Finally a unigram was selected only if it was not part of a common bigram to avoid duplicity. Then by applying the limit of 0.015 only 470 unigrams ended up in the final vocabulary.

Once the final vocabulary was selected (640 between unigrams and bigrams), the survey was formatted accordingly to use sIBP and split into training and validation sets (50%- 50%). Using the training, parameters are estimated via variational inference, in ten iterations to evaluate the output at several local models. The ideal hyperparameter configuration was selected based on how different the treatments were evidenced in the Coherence and Exclusivity Score (CE score). Ideally, the higher CE would be the best, however as the objective to evaluate specific hypothesis, the authors chose between top 3 configurations and selected the one that was closer to their initial hypothesis manually.

With the selected model, the authors applied it to the test set where a soft assignment (Z) was made for each input based on the vector words and parameters found in the selected model in training. To get a soft assignment they used a random Bernoulli distribution with input parameter z, resulting in a the matrix $\hat{Z}$. In order to get a relationship with the target variable (thermometer in this case) and measure the effect of the treatment, they used a linear regression as the coefficient will capture the relation. As doing this procedure once would not give necessarily a good estimate, they decided to do a 1,000 bootstrap procedure to infer the confidence intervals of the coefficients. A mean of these coefficients was considered to expose the causal effect on the responses, and it was plotted in a graph to ease their understanding.

3 Replication Results

The pre-processing part was done in Python, the replication code was rebuilt based on the original in an iPython notebook. All the encoding of the files was set to UTF-8 as the inputs could include words from other languages. We could understand the logic they used in the construction of the datasets. With the inputs the authors provided us (vocabulary, stopwords and survey) we were able to recreate their final 3 files successfully.

The modeling part was coded in R, all the variational inference was written in a R file to which we added a seed at the beginning of the document as the authors recommended. Then, we wrote a R markdown document with details in each step of the process. We selected the best configuration among top 6 CE score results having in mind the same hypothesis as Grimmer and Fong, finding what treatments based on professional background affect the thermometer.

Given that we set the seed in this replication and it was not set in the original run, the models vary a bit. Despite this, we could find results consistent to Grimmer and Fong. Tables 1 and 2 show the replication and original results correspondingly.

Table 1. Replication: Top 10 words in each treatment

Treatment1	Treatment2	Treatment3	Treatment4	Treatment5	Treatment6	Treatment7	Treatment8	Treatment9	Treatment10
delta	law	united_states	germany	served	united_states	high_school	fraternity	democratic	bachelor_science
fraternity	teacher	military	earned_jd	elected	served	political_science	immigrant	legislative	science_degree
theta	office	air_force	jewish	democratic	elected	board	maternal	office	father
phi	began	medal	maternal_grandparents	incumbent	infantry	juris_doctor	mother	republicans	mother
medal	court	enlisted	immigrants	republican	division	bachelor_arts	grandfather	election	southern_california
united_states	day	army	magna_cum	state	enlisted	office	née	alpha	business
kappa	age	states_army	air	seat	star	elected	irish	campaign	earned_bachelor
bronze	daughter	combat	combat	defeated	republican	president	kappa	republican	family
member	youth	air	maternal	election	distinguished	law	phi	appointed	security
vietnam	seat	service	public_school	air_force	army	political	epsilon	fraternity	help

Table 2. Original : Top Words for 10 Treatments sIBP Discovered

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

As can be seen the order of the topics has changed, but the groups still look alike. In the original paper there are two treatments related to army and war (Nm. 10 and 6) this groups are covered by groups 10 and 3. There is also, a law related treatment in Nm. 9 which has its homologous in group 7 in our replication. There are also, some treatments focused in elections,

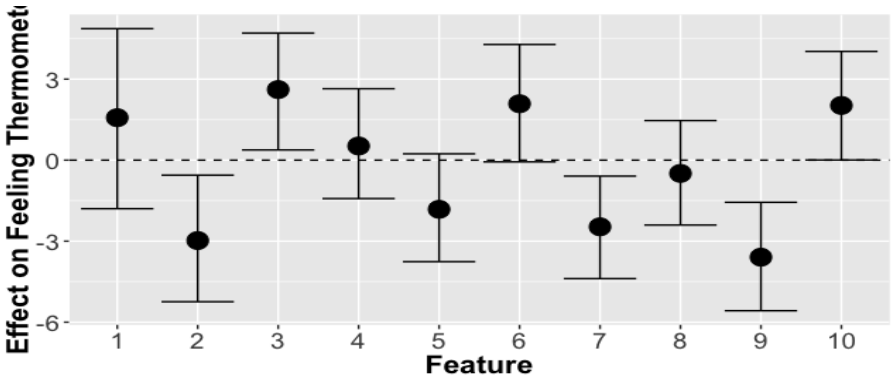

Figure 2. Replication - 95% Confidence Interval

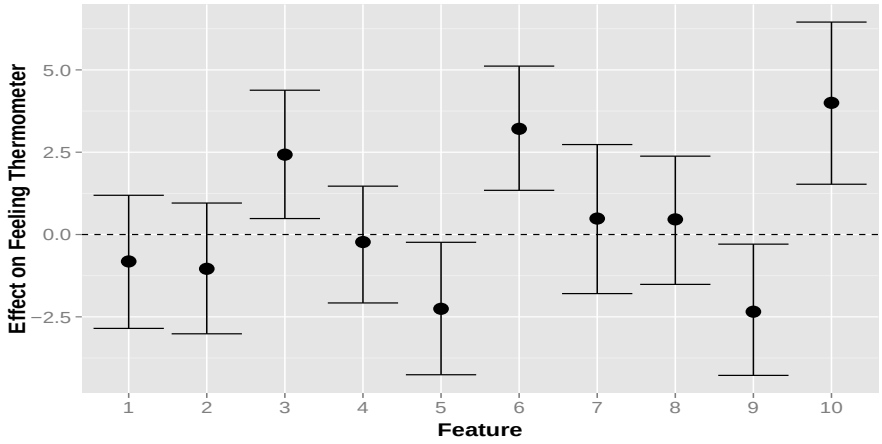

Figure 3. Original - 95% Confidence Interval

democrats and republicans in groups 5 and 7 originally, evidenced in groups 5 and 7 in the replication. Finally it is worth noticing that a mixed group that seems to have more family content (Nm. 1 in the original) is present in the replication too in groups 4 and 8. Hence, we can say that despite the order it is possible to see that the groups obtained were similar to the ones published in Fong et al.

With the words, now we could see how the causal effects affect the thermometer by using confidence interval graphs. Figure 2 and Figure 3 show the results obtained in the test set with the replication and the original setup.

Analyzing the results, we can see that originally treatments 5 and 9 negatively affected the response on the thermometer. These correspond to law school and election/democracy groups. Replication results say something similar as group 7 and 9 reveal a distaste for this background in politicians. As stated in Fong et al, it is shown that there is not aversion to education as treatment 10 in the original and treatment 3 in the replication affect positively. There is also a positive perception to military experience as in treatment 6 in the original and 3(and 6) in our replication.

4 Experiments with Yelp Reviews

4.1 Dataset and Approaches

We used the dataset provided by Yelp in the ”Academic Challenge”, which consists of around 2.7 million reviews in 10 cities around the world and gathers information of more than 86,000 businesses. Our target variable in this case is the rating (0-5) the customer gave a business, and our dependent text is the review.

We had two approaches in our experimental setting:

- All the dataset:

In this approach we used all the dataset which included reviews of hotels, restaurants, bars, home services and more. The range of variability was pretty large, but we wanted to see whether the method would capture this sort of relations.

- Specific dataset:
  Given that the previous approach would be too wide to be captured by the model, as suggestion of our adviser, we run the model with a shorter niche of reviews. We chose Nevada state in the US focused in restaurant reviews of Mexican food.

4.2 Differences with Fong et al.

It is worth noticing that our working framework differs in some aspects from the one mentioned in Fong et al. In the original work, authors collected information using a survey, providing the respondents with a pre processed text and gathering the ranking the person considers based on that document. They show that doing this ensures that the text is sufficient to identify treatment effects.

Some assumptions are also made, one of them is a version of the Stable Unit Treatment Value Assumption (SUTVA) which assures that each respondent treatment assignment depends only in her assigned text. Also, they assume that an individual would respond in the same way to two texts if they have the same latent feature.

In our experiment, the nature of Yelp reviews suggests that the text was already known by the respondent. In fact, the text was written by the respondent based on a previous experience of the place he visited. Despite this, we believe that the sIBP still holds to the experiments given that the review can have several treatments behind that are not necessarily implied in the text. We assume that these treatments are lead by the main characteristics a business shows toward customers such as service, quality of the good/food offered, ambiance, etc. We're considering this assumption makes the sIBP fit to our experimental dataset problem, and gives it more flexibility than a supervised LDA model, making our results still valid.

4.3 Experiment - All Reviews

There are exactly 2,685,066 reviews in the dataset, with such a size we decided to start our analysis on a random sample of 5,000 which had 3.7038 of average star measure. As our Y was standardized, the mean will be the 0 of the new standardized distribution.

Our preprocessing process was based on Fong et al., with a few modifications of our own. We used the same stopwords and excluding punctuation signs and numbers, with the inputs provided by Grimmer and Fong. Similarly to what they did in their experiment, we first select a common vocabulary, to be used in the final bigram and unigrams selection. As we did not know what threshold to use, we decided to try several ones, ending up with models to be labeled in the following sections as V600, V800 and V1000. Details can be found on Table 3.

Table 3. Yelp all reviews - Vocabulary size details

Threshold	Num.Unigrams	Num. Bigrams	Total Vocab.Size	Model Name
0.011	616	24	640	V600
0.009	758	36	794	V800
0.007	975	56	1031	V100

To have an initial idea of the results, we launched a 462 vocabulary size model, with 10 topics. We found them to be too mixed, and noticed that it was possible to group some of the treatments, identifying 5 main topics we would like to get: service, location, taste/experience, bad things of the place and one mixed group.

Then, we set the number of treatments to 5, and run the model for our 3 datasets varying in vocabulary size. The one that gave us clearer results was the "V600".

The top 10 words for each of the 5 treatments can be found in Table 4. It is possible to say that the first treatment is related to menus and food while the second is related to compliments about the flavor and place. It seems that the fourth treatment has some relation with the third, although the third one seems to have more negative words. And, the last one talks about hotels and the other businesses (besides restaurant) that are on Yelp. How do these groups relate to the number of stars customers marked? Figure 4 says that. The treatments are represented in the X-axis while the (average) Z-score of the effect on thermometer values

Table 4. Yelp's Top 10 words for 5 treatments - V600

X1	X2	X3	X4	X5
sauce	wonderful	told	time	hotel
chicken	food_delicious	asked	ordered	stay
cooked	grab	minutes	order	stayed
ordered	great_place	called	asked	pool
salad	recommend_place	time	table	staying
fried	curry	wait	people	rooms
dish	authentic	manager	told	car
cheese	food_great	left	experience	room
flavor	brunch	phone	wanted	floor
meal	crowd	work	place	buy

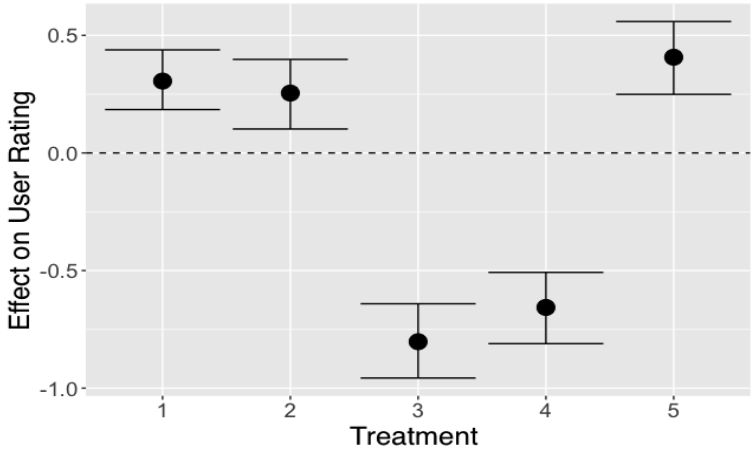

Figure 4. 95% Confidence Interval- Yelp V600 5 treatments

given by each treatment is on the Y-axis. As we mentioned before, the third and fourth treatments gave bad results. What is unexpected is that 3 ranks worse than 4. Treatment 5 resulted in a high rank, followed by the first two.

These results confirm our hypothesis. If the client had to wait for long or asked for the manager, it means that he had a bad experience hence will give a low rating. But, if he liked the food and place, as well as the service, a high rating is expected.

4.4 Mexican reviews in Nevada experiments

Based on our advisor's suggestion and given that the results shown in the previous section combined restaurants and hotels (mainly), we decided to experiment with a more focused dataset to have a selected vocabulary and to avoid huge discrepancies in opinions due to location. We decided to select Mexican food restaurants in Nevada.

The pre-processing part was similar to what we did before with the exception of the vocabulary size. In this case, we focused in a 1304 vocabulary size (threshold of .005) expecting to have a wider variety of topics as more n-grams will be available. The modeling part was launched considering 5 and 10 to test our hypothesis, for the same reason. We display results obtained with 10 topics, the other one can be found in the appendix.

The results obtained were intuitive too, but still confirmed that the causal effect of a certain feature of a restaurant will increase or decrease the number of starts it will receive from a customer.

As expected, taste has a positive effect in the stars of a business (treatment 2) as well as the sides and drinks the restaurant has to offer (treatment 7), both are way up the mean, generating an effect of 1.5 point increase and 0.8 in the standardized number of stars.

On the other hand, if the service is slow in a restaurant that will bring along bad comments and negative punctuation in Yelp reviews, mainly of complaints (treatments 5 and 9 as well as 10), each of them reduces around 0.5 points in the final star reviews.

Table 5. Yelp Mexican Top 10 words for 10 treatments

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
taco	table	meat	tacos_el	table	location	chips_salsa	bowl	worst	minutes
mexican	seated	cheese	el_gordo	place	business	appetizer	fresh	terrible	table
small	server	tortilla	lines	minutes	patient	great	fried_ice	attitude	food
choose	brought	taco	fries	asked	dine	dish	bean_dip	waste	asked
tacos_el	asked	tacos_el	meats	finally	told	bean_dip	great_experience	store	order
delicious	minutes	el_gordo	pollo	order	ago	sweet	chipotle	overcooked	offered
tacos	chips	tacos	sushi	night	reviews	mandalay	rice	girl	manager
carne_asada	water	carne_asada	taco_place	wait	sushi	margarita	pico_de	standing	water
el_gordo	manager	asked	bowls	ordered	management	tequila	de_gallo	food	left
lengua	empty	onions	best_tacos	half	read	delicious	seaweed	sat	time

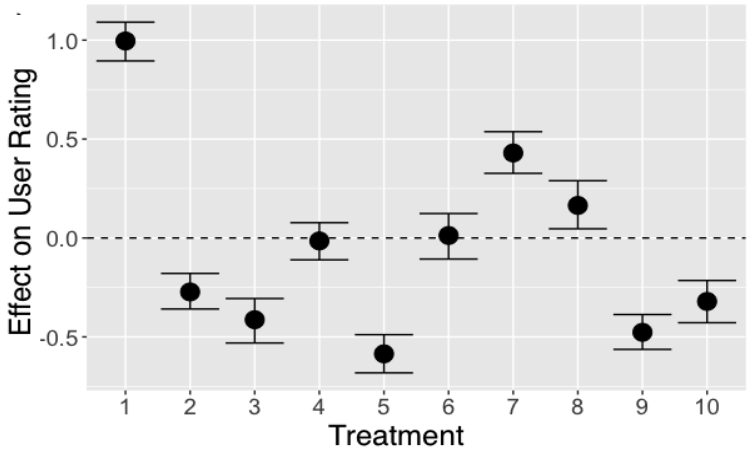

Figure 5. 95% Confidence Interval- Yelp Mexican selection, 10 treatments

5 Conclusions and Future Work

- The method proposed by Grimmer et al. can be applied to a myriad contexts of experimentation. One should be aware of the experimental settings.
- sIBP gives more flexibility in modeling the causal effects of contents on the user responses.
- Our experiments show that the methods works well in settings similar to the one proposed by Fong et al, as the treatments found were actually very intuitive.
- A bigger vocabulary seems to give a bigger chance to find more variability in the topics you may get. However, if the number of topics is set to a larger value and the number of real hidden topics is less than that, a big vocabulary will end up in mixed treatments.

Acknowledgements

We would like to thank Prof. David Sontag for guiding us throughout the project and helping us understand the material. We would also like to thank Dr Justin Grimmer and Christina Fong at Stanford University for helping us by sharing their code, data and valuable advice throughout the project.

Author contributions statement

Luisa worked on data preparation specifically and coordinated with the authors to ensure consistency in the approach. Varun worked on structuring the code and running the model for replication purposes and on Yelp reviews. Both the authors equally contributed towards understanding the methodology, designing and running the experiments and in writing up the report.

References

1. Fong, C. & Grimmer, J. Discovery of treatments from text corpora (2016).
2. Holland, P. W. Statistics and causal inference. *Journal of the American statistical Association* **81**, 945–960 (1986).
3. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66**, 688 (1974).
4. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* **3**, 993–1022 (2003).
5. Mcauliffe, J. D. & Blei, D. M. Supervised topic models. In *Advances in neural information processing systems*, 121–128 (2008).
6. Quadrianto, N., Sharmanska, V., Knowles, D. A. & Ghahramani, Z. The supervised ibp: Neighbourhood preserving infinite latent feature models. *arXiv preprint arXiv:1309.6858* (2013).
7. Doshi-Velez, F., Miller, K. T., Van Gael, J., Teh, Y. W. & Unit, G. Variational inference for the indian buffet process. In *Proceedings of the Intl. Conf. on Artificial Intelligence and Statistics*, vol. 12, 137–144 (2009).

Appendix

1. All Reviews: V800

Table 6. Top 10 words for 5 treatments - V800

X1	X2	X3	X4	X5
ordered	hotel	sauce	ordered	buy
table	casino	ordered	food	customer_service
order	rooms	chicken	menu	work
menu	floor	dish	place	shop
restaurant	stayed	flavor	restaurant	store
sauce	room	rice	sauce	care
tables	staying	cooked	side	car
good	stay	cheese	good	working
asked	nights	menu	time	help
food	bathroom	served	table	online

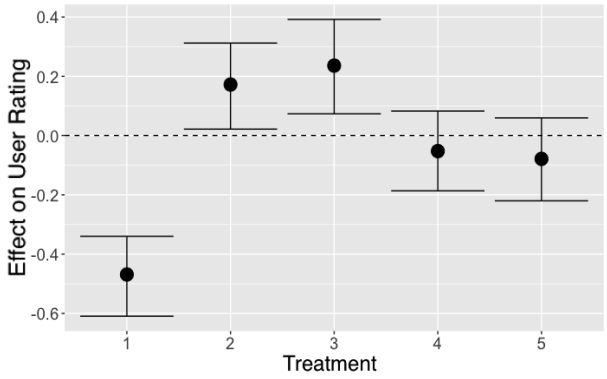

Figure 6. 95% Confidence Interval- Yelp V800,5 treatments

2. All Reviews: V1000

Table 7. Top 10 words for 5 treatments - V1000

X1	X2	X3	X4	X5
chicken	wide	knowledgeable	time	room
food	yum	shop	ordered	hotel
ordered	pizza	helped	order	stay
restaurant	cocktail	hang	food	called
menu	salads	highly_recommend	people	told
sauce	mexican	years	asked	call
dish	mac_cheese	fantastic	restaurant	work
cooked	filet	helpful	table	staying
bread	joint	positive	good	check
good	wings	salon	menu	front_desk

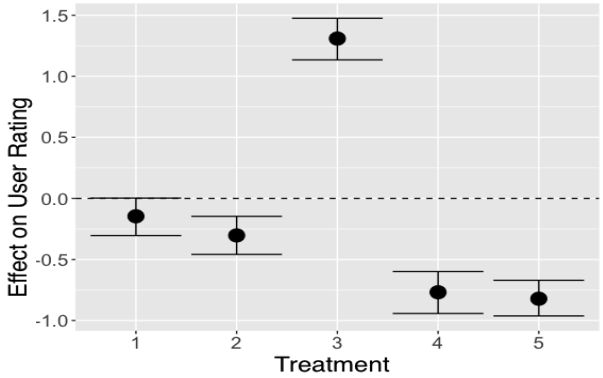

Figure 7. 95% Confidence Interval- Yelp V1000,5 treatments

3. Mexican reviews in Nevada: 5 topics

Table 8. Yelp Mexican Top 10 words for 5 treatments

X1	X2	X3	X4	X5
asked	great_food	told	taco	dish
order	awesome_food	minutes	tacos_el	rice
table	best_place	manager	el_gordo	menu
food	friendly_staff	service	meat	spicy
menu	excellent_service	waiting	tacos	good
ordered	best_mexican	rude	lengua	server
chips	friendly_helpful	worst	carne_asada	meal
side	good_food	awful	lines	refried_beans
good	amazing_food	waited_minutes	asada_fries	chips
restaurant	food_great	paid	adobada	chicken

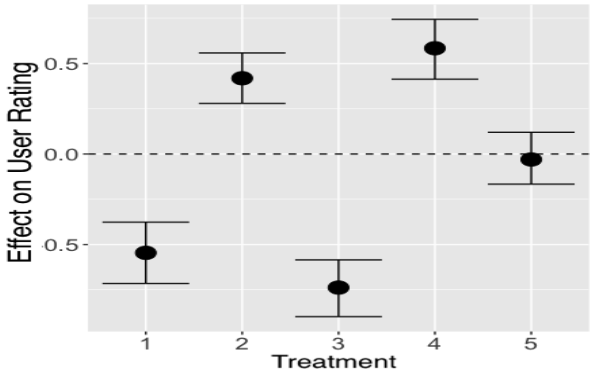

Figure 8. 95% Confidence Interval- Yelp Mexican selection, 5 treatments