Online Appendix 1: Cross-Validation Approach and Test data Performance For Binary Class Models

Cross Validation Approach

We randomly split the data into 70% training data and 30% test data. We estimate the learning rate and number of epochs of training via 5-fold cross-validation on the training data. We select the best parameters based on the 5 fold cross-validation results, and re-estimate the model on the entire training data set and report the performance on the test data. Table 5 contains these results. We varied learning rates, number of training epochs, loss-types, and model specifications. Within model specifications, we also estimated BERT+Linear, BERT + CNN and BERT + LSTM models. BERT+CNN is the best reported model for 'Pos_Clean'. In most instances the linear variant performed the best. Additionally, we account for the imbalanced nature of the data set by using dice loss and focal loss, in addition to the standard cross-entropy loss. We present the top 5 competitive models for each variable on the next 3 pages (based on the F1 Statistic on test data). In all we estimated 1,432 models.

Table A1: Top 5 Models for Each Attribute-Valence Combination (Based on F1 Statistic)

Attribute-Valence Combination	Learning_Rate	Loss Type/Variant	Model Name	Model Type	Train E...	F1 Test	Precision	Recall Test	ROC AUC	PR AUC
Discuss_Staff	5.00E-06	cross	xlnet-large-cased_cross	Xlnet	20	0.84	0.79	0.91	0.95	0.88
Discuss_Staff	5.00E-05	cross	xlnet-base-cased_cross	Xlnet	10	0.84	0.77	0.93	0.95	0.86
Discuss_Staff	3.00E-06	cross	xlnet-base-cased_cross	xlnet	40	0.83	0.76	0.93	0.95	0.90
Discuss_Staff	1.00E-05	cross	roberta-base_cross	roberta	50	0.83	0.81	0.85	0.93	0.89
Discuss_Staff	8.00E-06	cross	roberta-base_cross	roberta	50	0.83	0.79	0.87	0.94	0.87
Pos_Staff	1.00E-05	cross	xlnet-base-cased_cross	xlnet	10	0.82	0.75	0.91	0.96	0.88
Pos_Staff	3.00E-05	cross	xlnet-base-cased_cross	xlnet	10	0.82	0.76	0.88	0.94	0.86
Pos_Staff	5.00E-05	cross	xlnet-base-cased_cross	xlnet	10	0.80	0.81	0.79	0.95	0.77
Pos_Staff	8.00E-06	cross	roberta-base_cross	roberta	50	0.79	0.83	0.76	0.92	0.84
Pos_Staff	7.00E-06	cross	xlnet-base-cased_cross	xlnet	40	0.79	0.78	0.79	0.96	0.86
Neg_Staff	0.0001	cross	roberta-base_cross	roberta	20	0.79	0.77	0.82	0.92	0.58
Neg_Staff	7.00E-05	cross	roberta-base_cross	roberta	20	0.77	0.77	0.77	0.94	0.76
Neg_Staff	5.00E-05	cross	xlnet-base-cased_cross	xlnet	10	0.77	0.71	0.84	0.95	0.80
Neg_Staff	3.00E-05	cross	xlnet-base-cased_cross	xlnet	10	0.75	0.69	0.82	0.96	0.80
Neg_Staff	5.00E-05	cross	roberta-base_cross	roberta	50	0.75	0.79	0.70	0.94	0.83
Discuss_Wait	7.00E-06	cross	xlnet-large-cased_cross	xlnet	30	0.74	0.68	0.82	0.91	0.74
Discuss_Wait	3.00E-05	dice	roberta-base_dice	roberta	20	0.74	0.79	0.69	0.91	0.64
Discuss_Wait	5.00E-06	cross	xlnet-large-cased_cross	xlnet	30	0.74	0.69	0.79	0.90	0.70
Discuss_Wait	9.00E-06	cross	roberta-base_cross	roberta	40	0.73	0.71	0.74	0.93	0.71
Discuss_Wait	7.00E-06	cross	roberta-base_cross	roberta	40	0.72	0.72	0.72	0.92	0.67
Pos_Wait	7.00E-05	cross	roberta-base_cross	roberta	30	0.59	0.71	0.50	0.79	0.47
Pos_Wait	7.00E-06	cross	xlnet-large-cased_cross	xlnet	30	0.52	0.46	0.60	0.90	0.49
Pos_Wait	3.00E-05	dice	roberta-base_dice	roberta	20	0.50	0.67	0.40	0.83	0.46
Pos_Wait	7.00E-05	cross	xlnet-base-cased_cross	xlnet	20	0.48	0.37	0.70	0.80	0.42
Pos_Wait	3.00E-05	cross	xlnet-base-cased_cross	xlnet	20	0.48	0.40	0.60	0.84	0.47

Table A2: Top 5 Models for Each Attribute-Valence Combination (Based on F1 Statistic; contd.)

Attribute-Valence Combination	Learning_Rate	Loss Type/Variant	Model Name	Model Type	Train E...	F1 Test	Precisio	Recall Test	ROC AUC	PR AUC
Neg_Wait	5.00E-06	cross	xlnet-large-cased_cross	xlnet	30	0.73	0.80	0.67	0.93	0.71
Neg_Wait	7.00E-06	cross	xlnet-large-cased_cross	xlnet	30	0.73	0.80	0.67	0.89	0.66
Neg_Wait	3.00E-05	dice	roberta-base_dice	roberta	20	0.70	0.73	0.67	0.90	0.72
Neg_Wait	7.00E-06	cross	xlnet-base-cased_cross	xlnet	40	0.70	0.73	0.67	0.94	0.76
Neg_Wait	5.00E-06	dice	roberta-base_dice	roberta	40	0.69	0.61	0.79	0.92	0.65
Discuss_Clean	5.00E-05	cross	roberta-base_cross	roberta	20	0.53	0.48	0.58	0.89	0.31
Discuss_Clean	5.00E-05	dice	roberta-base_dice	roberta	20	0.53	0.48	0.58	0.89	0.31
Discuss_Clean	9.00E-05	cross	roberta-base_cross	roberta	20	0.53	0.48	0.58	0.72	0.39
Discuss_Clean	9.00E-05	dice	roberta-base_dice	roberta	20	0.53	0.48	0.58	0.72	0.39
Discuss_Clean	3.00E-05	cross	roberta-base_cross	roberta	20	0.52	0.42	0.67	0.87	0.34
Pos_Clean	1.00E-06	cnnbert	roberta-base_cnnbert	roberta	30	0.57	0.42	0.89	0.98	0.25
Pos_Clean	1.00E-06	cross	roberta-base_cross	roberta	30	0.57	0.42	0.89	0.98	0.25
Pos_Clean	1.00E-06	dice	roberta-base_dice	roberta	30	0.57	0.42	0.89	0.98	0.25
Pos_Clean	1.00E-06	focal	roberta-base_focal	roberta	30	0.57	0.42	0.89	0.98	0.25
Pos_Clean	5.00E-06	dice	roberta-base_dice	roberta	20	0.57	0.42	0.89	0.99	0.27
Neg_Clean	5.00E-05	cross	xlnet-base-cased_cross	xlnet	30	0.58	0.78	0.47	0.75	0.48
Neg_Clean	3.00E-05	cross	xlnet-base-cased_cross	xlnet	30	0.52	0.58	0.47	0.72	0.45
Neg_Clean	5.00E-06	dice	roberta-base_dice	roberta	20	0.48	0.60	0.40	0.78	0.32
Neg_Clean	1.00E-06	cnnbert	roberta-base_cnnbert	roberta	30	0.46	0.55	0.40	0.74	0.41
Neg_Clean	1.00E-06	cross	roberta-base_cross	roberta	20	0.46	0.55	0.40	0.74	0.41
Discuss_Price	9.00E-06	cross	xlnet-base-cased_cross	xlnet	50	0.73	0.68	0.79	0.90	0.64
Discuss_Price	5.00E-06	cross	roberta-large_cross	roberta	30	0.73	0.71	0.76	0.89	0.62
Discuss_Price	9.00E-06	cross	roberta-base_cross	roberta	30	0.73	0.72	0.74	0.90	0.68
Discuss_Price	5.00E-06	cross	roberta-base_cross	roberta	30	0.73	0.68	0.78	0.89	0.63
Discuss_Price	5.00E-05	cross	roberta-base_cross	roberta	50	0.72	0.66	0.81	0.90	0.58

Table A3: Top 5 Models for Each Attribute-Valence Combination (Based on F1 Statistic; contd.)

Attribute-Valence Combination	Learning_Rate	Loss Type/Variant	Model Name	Model Type	Train Epochs	F1 Test	Precision Test	Recall Test	ROC AUC Test	PR AUC Test
Pos_Price	7.00E-06	dice	roberta-base_dice	roberta	20	0.51	0.43	0.63	0.87	0.34
Pos_Price	0.0001	focal	roberta-base_focal	roberta	20	0.50	0.39	0.68	0.86	0.40
Pos_Price	1.00E-06	dice	roberta-base_dice	roberta	20	0.50	0.40	0.66	0.84	0.35
Pos_Price	1.00E-06	cross	roberta-base_cross	roberta	30	0.49	0.39	0.66	0.88	0.32
Pos_Price	9.00E-06	cross	roberta-large_cross	roberta	30	0.48	0.43	0.55	0.84	0.29
Neg_Price	7.00E-06	cross	roberta-base_cross	roberta	30	0.64	0.73	0.58	0.92	0.54
Neg_Price	5.00E-06	cross	roberta-large_cross	roberta	30	0.61	0.66	0.58	0.88	0.65
Neg_Price	9.00E-06	cross	xlnet-base-cased_cross	xlnet	50	0.61	0.74	0.52	0.89	0.55
Neg_Price	1.00E-05	cross	roberta-base_cross	roberta	20	0.58	0.79	0.45	0.92	0.62
Neg_Price	3.00E-05	dice	roberta-base_dice	roberta	20	0.56	0.82	0.42	0.93	0.65

Online Appendix 2: Model Comparison with Multiclass Models

We model each of three outcomes separately: incidence of positive discussion, incidence of negative discussion and incidence of discussion. We do not constrain the third outcome to equal the sum of the first two outcomes. Imposing this constraint leads to loss of predictive performance. Since maximizing predictive performance is the main objective of the text analysis, we prefer separate modelling of the three outcomes, to imposing the equality constraint.

We illustrate this below in the context of the "staff" attribute. We estimate 7 transformer models, each with 8 learning rates, leading to a total of 56 models. Each model has three possible outcome classes: positive discussion, negative discussion and no discussion, such that the probabilities of all three outcomes sums to 1. The data is split into training and test set. We test learning rates 5e-06, 1e-05, 2e-05, 5e-05, 8e-05, 1e-04, 5e-04, and 1e-03. We estimate the optimal learning rate and number of epochs using 3-fold cross-validation on the training dataset.

The table on the next page presents the test set performance of the best 25 best of these 56 models, sorted in descending order by Matthews Correlation Coefficient or MCC. MCC is a useful metric for evaluating multiclass models. The models were estimated on a 4 GPU server and estimated in parallel. We use 3-fold cross-validation to ensure adequate number of examples of each class are present in each fold. We use the F1 metric for choosing the best performing model. The F1 metric for the proposed model (see Table 4) which does not impose the equality constraint far exceeds the F1 metric of these models. For this reason, we retain the unrestricted model.

This result of the unrestricted model performing better than the multiclass model holds for all attributes. This might appear surprising at first. However, there are likely two reasons for this. First, each model has been pre-trained on very large corpora (running into gigabytes in a domain-agnostic manner) to learn linguistic relationships. Additionally, each model captures complex linguistic relationships via millions of parameters (the smallest model has 17 million parameters). Consequently, when we fine-tune these models on our dataset, across all parameters, the multiclass specification acts as a constraint on the optimal weights for any specific valence-attribute combination. This is exacerbated by the asymmetry in the labels for positive, negative and no-discussion, with no-discussion constituting the majority class by a large margin (i.e. the probability of a positive discussion or a negative discussion of an attribute, is much smaller than that of no discussion).

Table A4: Test Set Performance of Multiclass Models for the Staff Attribute

		Positive Staff			Negative Staff			
Model	**Learning Rate**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**MCC**
xlm-roberta-base	5.00E-05	0.74	0.71	0.76	0.77	0.81	0.74	0.72
allenai/longformer-base-4096	1.00E-05	0.72	0.77	0.67	0.74	0.79	0.69	0.71
allenai/longformer-base-4096	2.00E-05	0.70	0.77	0.65	0.75	0.78	0.72	0.71
roberta-base	2.00E-05	0.68	0.69	0.67	0.77	0.82	0.72	0.70
allenai/longformer-base-4096	5.00E-05	0.69	0.75	0.65	0.75	0.76	0.74	0.69
allenai/longformer-base-4096	5.00E-06	0.67	0.76	0.61	0.74	0.79	0.69	0.69
roberta-base	5.00E-05	0.69	0.65	0.73	0.75	0.76	0.74	0.69
bert-base-cased	1.00E-04	0.71	0.86	0.61	0.70	0.68	0.72	0.69
xlm-roberta-base	5.00E-06	0.67	0.71	0.63	0.67	0.73	0.62	0.66
allenai/longformer-base-4096	1.00E-04	0.69	0.76	0.63	0.65	0.81	0.54	0.65
roberta-base	5.00E-06	0.65	0.64	0.67	0.73	0.81	0.67	0.65
google/electra-base-discriminator	5.00E-05	0.65	0.65	0.65	0.71	0.70	0.72	0.64
roberta-base	1.00E-05	0.65	0.70	0.61	0.68	0.79	0.59	0.64
flaubert/flaubert_base_cased	1.00E-04	0.67	0.73	0.63	0.61	0.62	0.59	0.63
xlm-roberta-base	1.00E-05	0.66	0.72	0.61	0.65	0.72	0.59	0.63
roberta-base	8.00E-05	0.62	0.65	0.59	0.68	0.79	0.59	0.63
xlm-roberta-base	2.00E-05	0.66	0.72	0.61	0.64	0.67	0.62	0.63
google/electra-base-discriminator	1.00E-04	0.63	0.62	0.65	0.66	0.79	0.56	0.63
allenai/longformer-base-4096	8.00E-05	0.66	0.75	0.59	0.58	0.73	0.49	0.61
flaubert/flaubert_base_cased	2.00E-05	0.60	0.67	0.55	0.60	0.65	0.56	0.61
albert-base-v1	5.00E-05	0.57	0.57	0.57	0.58	0.67	0.51	0.60
flaubert/flaubert_base_cased	5.00E-05	0.61	0.62	0.61	0.57	0.58	0.56	0.60
bert-base-cased	8.00E-05	0.63	0.77	0.53	0.60	0.79	0.49	0.59
flaubert/flaubert_base_cased	1.00E-05	0.57	0.63	0.51	0.53	0.62	0.46	0.59
albert-base-v1	1.00E-04	0.58	0.61	0.55	0.60	0.71	0.51	0.59
google/electra-base-discriminator	8.00E-05	0.56	0.69	0.47	0.63	0.68	0.59	0.58

Online Appendix 3. Evidence from 9,000 simulated datasets of identification of the generalized synthetic control model

We estimated the model 9,000 times on 9,000 simulated datasets. We now describe this simulation. First we created simulated datasets based on the following equation:

$$y_{it} = constant + \alpha_i + \gamma_t + \lambda_i \times F_t + \delta_{it} + 1 \times Rvwrs - 1 \times Rvws + 5 \times Rstrnts - 2 \times trnd$$

where,
$constant$ = 0.5;
α_i : is a unit specific fixed effect drawn from uniform(-5,5);
γ_t : is a time specific fixed effect drawn from uniform(-5,5);
Rvwrs, Rvws, Rstrnts and *trnd* refer to number of reviewers, number of reviews, number of restaurants and linear time trend respectively. These variables are from our dataset, and are not simulated.
δ_{it} : is drawn from N(ATT, 1), where ATT takes values -3, -1 and 1 and 3 across different simulations.
λ_i : is the unit specific factor of size r, drawn from uniform(-0.06, 0.04) for control units and uniform(-0.08 ,0.23) for treatment units. The ranges were selected based on the range of values observed in the reported estimates.
F_t: is the time specific factor of size r, drawn from uniform(-2.93, 2.93). The range was selected based on the range of values observed in the reported estimates.
r is the number of factors. We assume 5 values: 1, 3, 5, 7 and 9.

We generated 1000 data sets for each combination of ATTs (4 values) and number of factors (5 values), resulting in 9,000 data sets. For each dataset, we first estimated the ATT, and then averaged the results across the 1000 datasets for each combination of ATT and r. Results, which appear below show, perhaps unsurprisingly, that the synthetic control approach is able to recover the ATT quite well across a wide range of ATTs and various number of factors.

Table A5: Means (and SE) of Estimates of ATTs Across 9,000 Datasets for Different Number of Factors

Number of factors (r)	True Value of ATT			
	-3	-1	1	3
1	-2.977 (0.328)	-0.992 (0.169)	1.008 (0.169)	3.023 (0.328)
3	-2.987 (0.354)	-0.977 (0.184)	1.023 (0.184)	3.013 (0.354)
5	-2.994 (0.378)	-0.988 (0.207)	1.012 (0.207)	3.006 (0.378)
7	-3.010 (0.372)	-1.004 ((0.199)	0.994 (0.199)	2.988 (0.373)
9	-3.051 (0.369)	-1.038 (0.199)	0.962 (0.199)	2.949 (0.369)

Online Appendix 4. Using pricing data to assess if treatment effects vary across restaurants with different price levels

Since we do not have data on prices in 2013, we use the following proxy for prices- we scrape the "$" measure of restaurant price from the webpage of each restaurant from the leading restaurant review website. This measure is on a 4-point ordinal scale (1, 2, 3 and 4; 1 being the lowest price), reflects the price for a meal per person, is culled by the website from recent consumer reviews of that restaurant, is computed at a restaurant level and not at a review level. Also, these prices are scraped in September 2020, aggregated from reviews over time, and not prices at the time of policy change. These indicators are also available for restaurants in our sample that have exited. The mean (SD) price for chains is 1.63 (0.52), and that for independents is 1.83 (0.48), suggesting that consumers perceive lower prices for independent restaurants. 51.1% of restaurants were rated 1, 46.5% of restaurants were rated 2, and the remaining 1.9% were rated 3 or 4.

To assess if the effect of the regulation varies by restaurant prices, we estimated a triple-difference (difference-in-difference-in-differences) regression model for independent restaurants[1]. This model extends the more common difference-in-differences specification to enable the assessment of heterogeneity in treatment effects.

$$y_{it} = Const_i + \beta_1 SanJose_i + \beta_2 Post_t + \beta_3 Price_i + \beta_4 SanJose_i \times Post_t + \beta_5 SanJose_i \times Price_i + \beta_6 SanJose_i \times Price_i \times Post_t + \beta_7 Rating_{it} + \epsilon_{it}$$

The subscript *i* indexes restaurants and *t* indexes time, in months. y_{it} is the mean (across all reviews of restaurant i posted in time t) of probability scores of discussion of an attribute-valence combination, obtained from the text model. $Post_t$ is 1 if the observation is from the post-treatment period, 0 otherwise. $SanJose_i$ is the dummy variable which accounts for the effect of belonging to the treatment group. $Price_i$ is the price level of restaurant i. The interaction term of interest is $SanJose_i \times Price_i \times Post_t$. It is unclear what we might predict about the sign of β_6. Nonetheless, a significant coefficient might give us insights on potential wage dynamics, or wage-price passthroughs. We estimate this model separately for each attribute-valence combination, for a 12-month window. Estimates of β_6 appear in the table on the next page. Addition of further controls (such as review length and the interaction term $Price_i \times Post_t$), and estimating the model for different time windows, does not alter this result.

[1] In the current synthetic controls specification in the paper, we construct synthetic controls at the zipcode-restaurant_type-month level. We do not have sufficient variation in the data to construct valid synthetic controls at the zipcode-restaurant_type-price_type-month level. So we prefer a triple difference specification to a synthetic controls specification. Under certain conditions, the triple difference estimator does not require the parallel trends assumption for causal inference (Olden 2018).

Table A6: Estimates of Coefficient of $SanJose_i \times Price_i \times Post_t$ for Triple-Difference Model for Independent Restaurants (Mean and SE)

Attribute-Valence Combination	Coefficient Estimates of $SanJose_i \times Price_i \times Post_t$
Discuss_Staff	-0.011 (0.021)
Pos_Staff	-0.013 (0.023)
Neg_Staff	-0.016 (0.015)
Discuss_Wait	0.000 (0.017)
Pos_Wait	0.004 (0.011)
Neg_Wait	0.016 (0.011)
Discuss_Clean	0.006 (0.011)
Pos_Clean	0.005 (0.011)
Neg_Clean	-0.005 (0.008)
Discuss_Price	-0.009 (0.022)
Pos_Price	-0.016 (0.019)
Neg_Price	-0.005 (0.010)

Note 1: $p > 0.05$ for all estimates

Note 2: Discuss_Staff measures whether the courtesy and friendliness of workers was discussed (1) or not (0); Pos_Staff measures whether the courtesy and friendliness of workers was discussed positively (1) or not (0); Neg_Staff measures whether the courtesy and friendliness of workers was discussed negatively (1) or not (0); Discuss_Wait measures whether waiting time (for seating, receiving food and check) was discussed (1) or not (0); Discuss_Clean measures whether cleanliness of restaurants was discussed (1) or not (0); Discuss_Price measures whether price of food or beverage items was discussed (1) or not (0). Other attribute_valence combinations follow along similar lines.

This coefficient is not statistically significant for any attribute-valence combination. This could be because of price data noisiness, and the wage-price pass-through noisiness. That is, our data might be insufficient for robust estimation of heterogeneity in treatment effects, across restaurants with different pre-treatment wages. Conditional on data quality, we do not find evidence of differences in treatment effects across restaurants with different price levels. Similar results hold for chains.

To ensure the results are not an artifact of small numbers in price classes, we aggregated the two lower price-tiered restaurants and the 2 higher price-tiered restaurants, and re-ran the above analysis. The results are unchanged.