# Natural Language Texts for Causal Inference in Randomized Experiments

Alex Xiaotong Gui[1], advised by Prof. Michael Baiocchi[2]

[1] *Department of Statistics,* [2]*Stanford School of Medicine*

## Overview

**Problem statement**

Can we let participants in a randomized trial answer open-ended questions and quantify the treatment-control differences using free responses? [1]

**Impact**

- less confirmation bias and thus more accurate insights
- work with effects that can't be measured by multiple choice

**Our work**

We proposed a testing procedure using supervised learning models and NLP framework to evaluate the causal effect of a treatment as measured by free-texts.[2]

**Major accomplishments**

- Built a classifier with features extracted from open responses to predict if a participant is in treatment or control
- Use the classifier's prediction accuracy as the statistic, do permutation test to measure the causal effect

## Data

**Figure 1** *The house participants are asked to describe.*

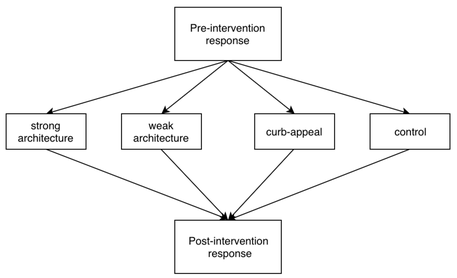


**Figure 2** *A workflow of the experiment set up.*

- Pilot survey on Amazon Mechanical Turks
- Three levels of interventions, one control group
- 100 observations in each intervention/control arm
- 41582 words in total, average 221 words per response
- Challenge: small data set

## Feature Extraction

**Preprocessing**

Performed normalization, removed contractions, and converted all words to lower case.

**Feature extraction methods**

- Bag-of-words
  - count token frequency and produce a mapping dictionary
  - removed all stop words
- Self-trained Word2Vec embedding [3]
- Pre-trained GloVe on Wikipedia and Gigaword, dim = 50 [4]
- Pre-trained GloVe on Wikipedia and Gigaword, dim = 100

## Model Selection

**Baseline** Bernoulli and Multinomial Naive Bayes

- Most indicative words metric:

$$log\left(\frac{P(word\_i \,|\, treatment)}{P(word\_i \,|\, control)}\right)$$

- Top indicative words between arch_strong and control returned by Multinomial NB:
  georgian, symmetrical, bay, pane, gabled, symmetry

**Experiments**

- Logistic regression with Word2Vec and GloVe embeddings
- SVM with Word2Vec and GloVe embeddings
- A two-layer dense neural network with ReLU and softmax activation

We split 20% data to test set and used 5-fold cross validation accuracy for model selection except for the neural network model. Since the classes are well-balanced, F1 score is almost identical with accuracy.

| Model | 5-fold cv | test |
|---|---|---|
| bern_nb | 0.7867 | 0.7895 |
| multi_nb | 0.8067 | 0.8158 |
| logreg_w2v | 0.48 | 0.6053 |
| logreg_glove_big | 0.8067 | 0.8421 |
| logreg_glove_small | 0.7867 | 0.8158 |
| svm_w2v | 0.44 | 0.5263 |
| svm_glove_big | 0.7267 | 0.7358 |
| svm_glove_small | 0.7133 | 0.7133 |
| neural net with TF-IDF | 0.8333 (dev) | 0.8421 |

## Permutation Test Results

We chose logistic regression with pre-trained GloVe embedding as the optimal classifier since it has good predictive performance and is very fast to train, which is a great advantage for permutation test.

We use the training accuracy of our classifier on current data as our test statistic. To create the null distribution, we randomly permute the class label, fit the classifier again and record the accuracy. The intuition is if the treatment has no effects, then the classifier can't predict treatment v.s control well and randomly switching the labels would make no difference in accuracy.

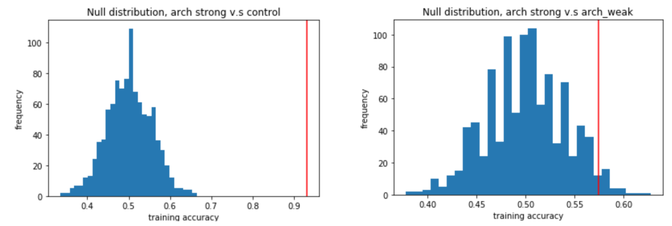


**Figure 3:** ***left.*** *Null distribution of arch_strong v.s control group. The red line is the test statistic. p-value = 0* ***B.*** *Null distribution of arch_strong v.s arch_weak. p-value = 0.027*

## Discussions and Future Directions

### Discussions

- Our dataset is incredibly small compared to conventional machine learning tasks, so simpler models tend to perform better than e.g. deep neural nets.
- Due to the same problem, training word2vec embedding on our own data gave terrible performance. However, using pre-trained GloVe embedding significantly improves the accuracy. We also deliberately chose Wikipedia data since it matches our research context (compared to Twitter data)
- Note that the end goal of this project is not to make predictions but to devise causal inference mechanisms.
- The final result suggests that people do describe the house differently when they're exposed to different education materials.

### Future work:

- Explore universal word embedding such as ELMo.
- Collect more training samples and in the mean time read more on methods to address small dataset.
- Evaluate how randomization affects testing results.

## References

[1] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2018:4586–4598, 2018.

[2] Liuyi Yao, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao, and Aidong Zhang. On the estimation of treatment effect with text covariates. In *IJCAI*, 2019.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[4] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.