Does double-blind review policy matter?

Findings from leading machine learning conference shows a small and relevant decrease in review score.

INTRODUCTION

- Peer Review is a cornerstone of scientific evaluation:
  - Educational institutional rankings
  - Production of graduate curriculum
  - Millions of dollars worth of grant funding each year
- Causal Inference with Text is an active area research

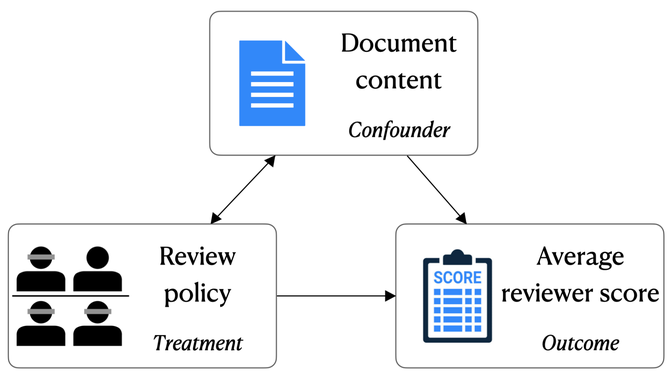
EXPERIMENTAL SETUP

$$\tau^{ATC} = \frac{1}{N_{T_0}} \sum_{i \in T_0} \Big(Y_i(T_i = 1) - Y_i(T_i = 0) \Big)$$

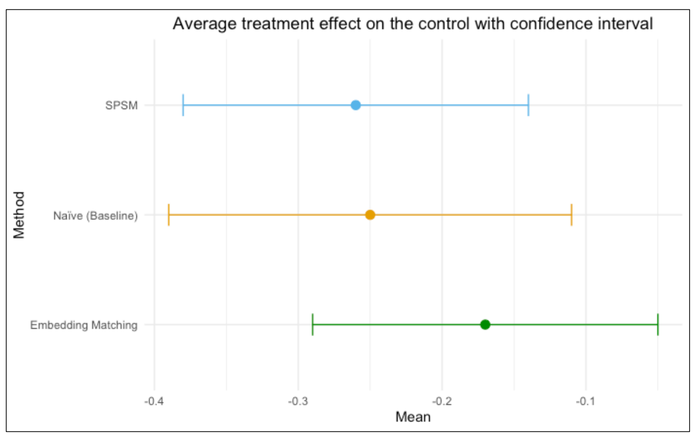
METHODS

1. Potential Outcome Matching Estimator
2. SPECTER Word Embedding
3. KNN-Matching

Raymond Zhang | Advisor Dan McFarland
Submitted to Annual Meeting of Association for Computational Linguistics (ACL)



RESULTS

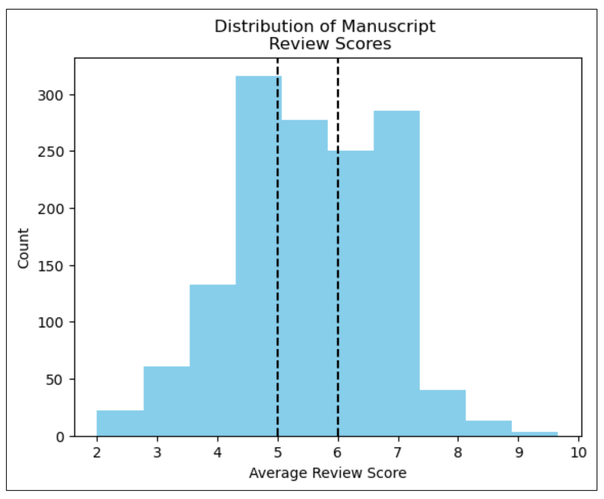


- ATU is -0.17 with 95% confidence interval [-0.29,-0.05]
- Human evaluation shows 71% prefer our method

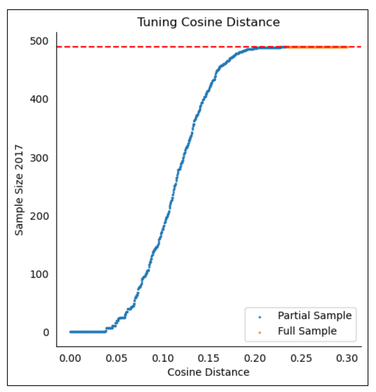
DISCUSSION/LIMITATIONS

- Unobserved confounders such as demographic attributes and institutional prestige
- **Theory of Change: Deindividuation**

Additional Information



Conference Year	Data Size
ICLR 2017	490
ICLR 2018	910



$$\hat{Y}_i(1) = \frac{1}{|M_i|} \sum_{j \in M_i} Y_j$$

$$\hat{\tau}^{ATT}_{\text{match}} = \frac{1}{N_{T_0}} \sum_{i \in T_0} \Big(\hat{Y}_i(1) - Y_i \Big)$$

Potential Outcome Matching Estimator

Stanford GRADUATE SCHOOL OF EDUCATION

LELAND STANFORD JUNIOR UNIVERSITY · DIE LUFT DER FREIHEIT WEHT · 1891