Feature Selection as Causal Inference: Experiments with Text Classification

College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

Michael J. Paul • mpaul@colorado.edu

University of Colorado Boulder

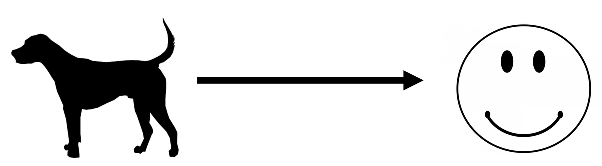
INTRODUCTION

Classifiers learn associations between features and classes. For a variety of reasons, these associations can be noisy and misleading.

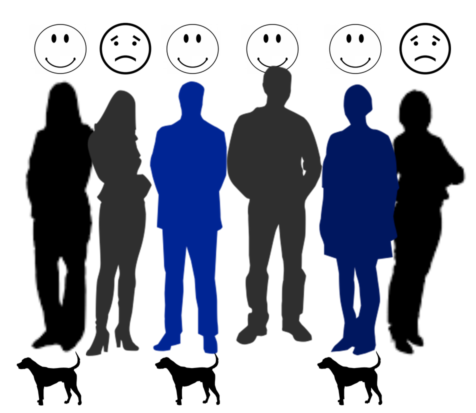
This work uses causal inference methods to learn more accurate feature associations. One goal of presenting these ideas is to generate new ideas for how to incorporate these techniques into NLP methods.

CAUSALITY AND PROPENSITY SCORE MATCHING

Suppose you want to test a hypothesis:
Getting a dog will make you happier.

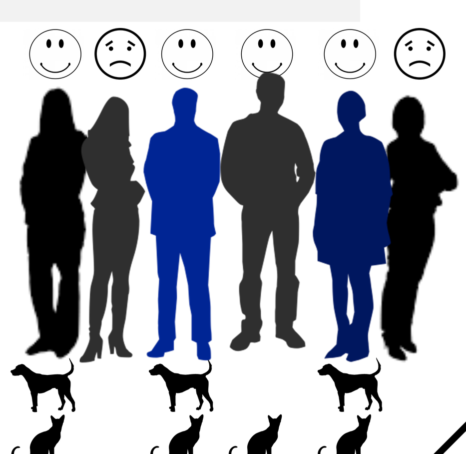


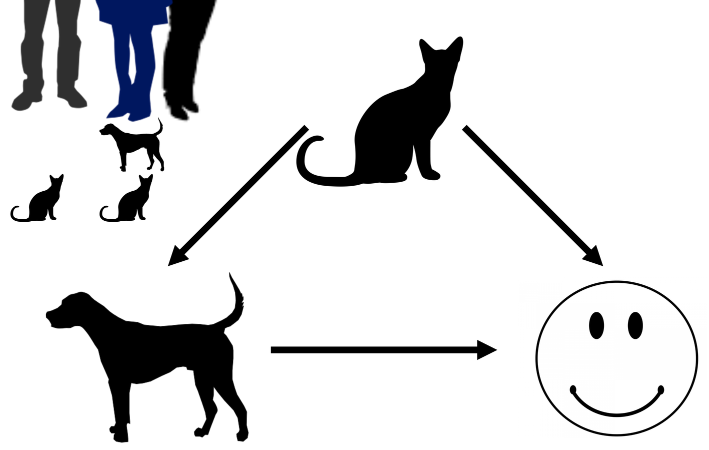
You might approach this by randomly sampling people, then measuring their current happiness and whether they own a dog.



The association between dogs and happiness could be misleading.

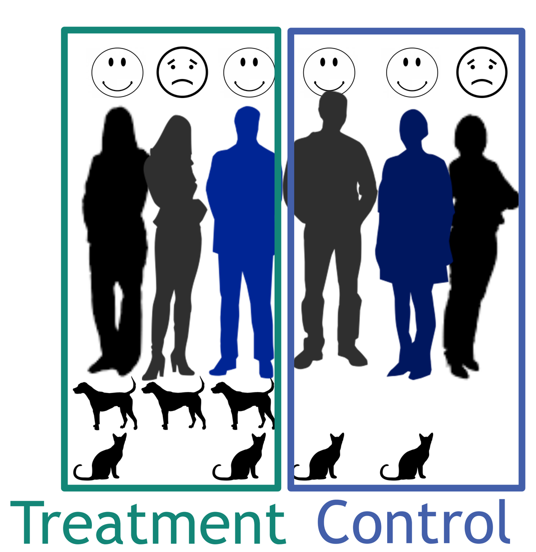
Maybe cats increase happiness, and cat owners are more likely to own dogs.





Cat ownership is third variable that interacts with both dog ownership and happiness, called a **confounding** variable.

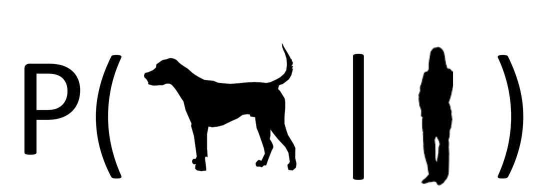
A **randomized controlled trial** randomly assigns **subjects** to receive the **treatment** (dog ownership), and then compares the **outcomes** (happiness) of people who did or did not receive treatment.



What if random assignment isn't possible?

One way to simulate the assignment to treatment vs control groups is to **match** individuals who are similar except in whether they had treatment [1].

One metric for matching people is their **propensity score**: the probability of receiving treatment [2].



The goal of matching people who are similarly likely to have received treatment (e.g., owning a dog) is that matched subjects will have the same distribution of other attributes (e.g., owning a cat), so that any difference in outcomes is likely due to difference in treatment alone.

Statistical tests can be used to determine differences in outcomes.

LEARNING BETTER WORD-CLASS ASSOCIATIONS

Idea: which word features *cause* a document to have the label that it has?

PEOPLE	TEXT
Subject	Document
Treatment	Word
Outcome	Class label

Recent work has shown that there are a number of sources of confounding bias in text classification [3].

We can formulate this as a traditional causal experiment, using propensity score matching to match documents that do and do not contain a word feature.

For each feature:

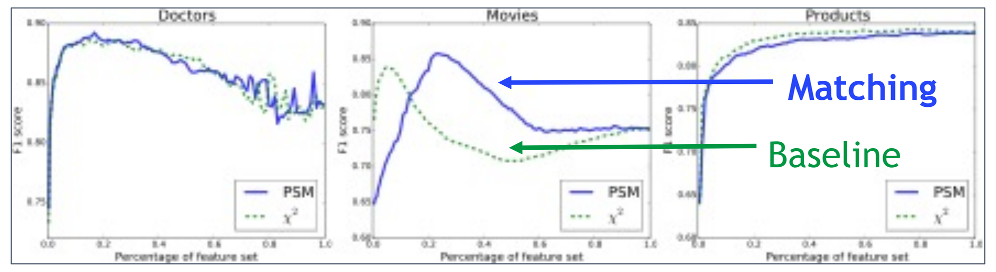
1. Define and calculate propensity scores
- Each document's probability of containing a word
- Logistic regression model using all other features

2. Match documents with similar propensity scores
- There are many variations of matching [4]
- This work used greedy one-to-one matching
  (one document that contains the feature with one that does not)

3. Calculate significance of feature
- McNemar's test statistic for chi-squared distribution [5]
  (similar to standard chi-squared test, but for paired data)

EXPERIMENTS

Document classification:
- Binary sentiment classes
- Bag of words features
- 3 review datasets from 3 domains
- Baseline: chi-squared test



F1-score (y-axis) when using only *n*% (x-axis) of the features, ranked by significance

Doctors		*Movies*		*Products*	
PSM	χ^2	PSM	χ^2	PSM	χ^2
great	told	great	worst	excellent	waste
caring	great	excellent	bad	wonderful	money
rude	rude	wonderful	and	great	great
best	best	best	great	waste	worst
excellent	said	love	waste	bad	best

Most significant features for sentiment classification

Training Corpus	Test Corpus					
	Doctors		*Movies*		*Products*	
	PSM	χ^2	PSM	χ^2	PSM	χ^2
Doctors	**.8569**	.8560	**.6796**	.6657	**.6670**	.6367
Movies	**.6510**	.5497	**.8094**	.7421	**.6658**	.4917
Products	.7799	**.7853**	**.8299**	.8245	.8234	**.8277**

Area under the feature selection curves (above)

When does it work?

The proposed method gives the largest gains when:
- testing on different domains
  - potential for better *generalizability*?
- using only a few features
  - potential for better *interpretability*?

Why does it work?

Example: *said*

Treatment (Propensity score: .80)

She repeatedly said, "I don't care how you feel" when my wife told her the medication (birth control) was causing issues. She failed to mention a positive test result, giving a clean bill of health.

Control (Propensity score: .79)

After a long, long conversation during which I tried to explain that I did not have records as I was only looked at by a sport trainer, they still would not see me without previous records.

CHALLENGES

- Scalability: this work required training a logistic regression model and performing document matching for every feature.
- Other ways to define the propensity score, or other general purpose metrics to use for matching?
- Other ways to incorporate propensity score matching into document classification beyond feature selection?
- How to use these ideas with dense feature representations?

RELATED WORK

Matching:
- Matching in NLP [6-8]
- Propensity score matching for text [9]
- Contrastive estimation [10]

Feature importance:
- Feature labeling [11,12]
- Annotator rationales [13]

REFERENCES

[1] P.R. Rosenbaum, D.B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39:33-38.

[2] P.R. Rosenbaum, D.B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41-55.

[3] V. Landeiro, A. Culotta. 2016. Robust text classification in the presence of confounding bias. *AAAI*.

[4] P.C. Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 46(3):399-424.

[5] Q. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2).

[6] U. Pavalanathan, J. Eisenstein. 2016. Emoticons vs. emojis on Twitter: A causal inference approach. *AAAI Spring Symposium*.

[7] C. Tan, L. Lee, B. Pang. 2014. The effect of wording on message propagation: topic- and author-controlled experiments on Twitter. *ACL*.

[8] Y. Zhang, E. Willis, M.J. Paul, N. Elhadad, B.C. Wallace. 2016. Characterizing the (perceived) newsworthiness of health science articles: A data-driven approach. *JMIR Med Inform* 4(3):e27.

[9] M. De Choudhury, E. Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. *ICWSM*.

[10] N.A. Smith, J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. *ACL*.

[11] H. Raghavan, O. Madani, R. Jones. 2006. Active learning with feedback on features and instances. J. Mach. Learn. Res. 7:1655-1686.

[12] G. Druck, B. Settles, A. McCallum. 2009. Active learning by labeling features. *EMNLP*.

[13] O.F. Zaidan, J. Eisner, C. Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. *NAACL*.