ROBUST TEXT CLASSIFICATION IN THE PRESENCE OF CONFOUNDING BIAS

Virgile Landeiro & Aron Culotta
Illinois Institute of Technology
Chicago

INTRODUCTION

- Development of text classification over more than 50 years[1]
- Mostly centered around categorization of documents into topics
- New areas of research (computational science):
  - Public health surveillance
  - Political science
  - Marketing
  - etc
- But algorithms stay the same: standard supervised classification algorithms
- To ensure validity of study ➔ need classifiers robust to confounding variables

1 - Maron, M. E. 1961. Automatic indexing: an experimental inquiry. Journal of the ACM (JACM) 8(3):404–417.

nyc	angeles	ny	york	california
los	la	brooklyn	snow	disneyland
jersey	city	san	ca	hollywood
monica	santa	nj	manhattan	losangeles
earthquake	team	dodgers	hills	cute
heart	vegas	chill	state	happiness
makeup	pacific	cali	father	brother
also	guess	socal	field	job
cant	venice	tacos	boo	wonderful
laugh	train	single	wanna	brothers

50 TOP FEATURES FOR LOGISTIC REGRESSION

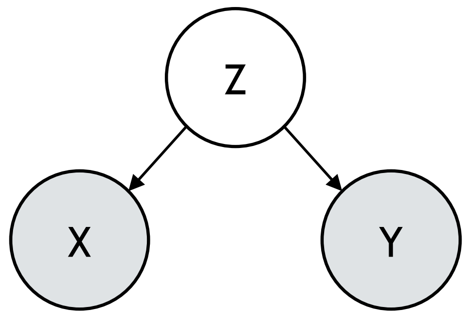
Female (resp. Male) and New York (resp. Los Angeles) are highly correlated.

nyc	angeles	ny	york	california
los	la	brooklyn	snow	disneyland
jersey	city	san	ca	hollywood
monica	santa	nj	manhattan	losangeles
earthquake	team	dodgers	hills	**cute**
heart	vegas	chill	state	**happiness**
makeup	pacific	cali	father	brother
also	guess	socal	field	job
cant	venice	tacos	**boo**	wonderful
laugh	train	single	wanna	brothers

50 TOP FEATURES FOR LOGISTIC REGRESSION

Male (resp. Female) and New York (resp. Los Angeles) are highly correlated.

WHAT IS A CONFOUNDING VARIABLE?



Graphical model: a confounding variable Z correlated with both X and Y.

- Prediction vs. causal inference.
- Assume same impact in training and testing sets.
- Small training datasets;
- Confounder shifts over time.

RELATED WORK

- Social science:
  - Matching
  - Stratification
  - **J. Pearl developed the back-door adjustment**
- Machine learning:
  - Selection bias[1]:

$$P_{train}(X) \neq P_{test}(X)$$

  - Changing target distribution:

$$P_{train}(Y) \neq P_{test}(Y)$$

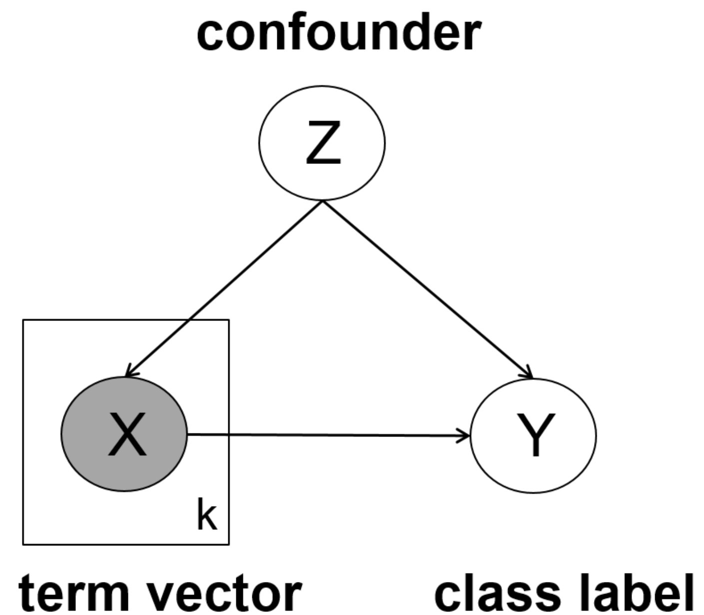
- We focus on:

$$P_{train}(Y|Z) \neq P_{test}(Y|Z)$$

1 - Bareinboim, Elias and Judea Pearl. "Controlling Selection Bias in Causal Inference." *AAAI*(2011).

BACK-DOOR ADJUSTMENT FOR TEXT CLASSIFICATION

- $D = \{(\boldsymbol{x_i}, y_i, z_i)\}_{i=1}^{n}$
- The back-door criterion requires that:
  - No node in Z is a descendant of X;
  - Z blocks every path between X and Y that contains an arrow pointing to X.
- The back-door criterion is met:



$$p(y|do(\boldsymbol{x})) = \sum_{z \in Z} p(y|\boldsymbol{x}, z) \times p(z)$$

BACK-DOOR ADJUSTMENT FOR TEXT CLASSIFICATION

$$p(y|do(\boldsymbol{x})) = \sum_{z \in Z} p(y|\boldsymbol{x}, z) \times p(z)$$

- Restrict to binary variables.
- Fit a logistic regression model on $p(y|\boldsymbol{x}, z)$ at training time by appending two features $c_{i,0}$ and $c_{i,1}$ to every $\boldsymbol{x_i}$.
- z is not observed at testing time.

Features matrix

$\boldsymbol{x_0}$	$\boldsymbol{x_1}$	$\boldsymbol{c_0}$	$\boldsymbol{c_1}$
0	0	0	1
0	1	0	1
1	0	1	0
1	1	0	1

$\boldsymbol{z}$
1
1
0
1

Dataset	Target variable	Confounder
Twitter	Location of a user: New York City or Los Angeles	Gender of the user: Male or Female
IMDb	Sentiment of the review: Positive or Negative	Genre of the film: Horror or Other
Canadian Parliament	Political affiliation: Liberal or Conservative	Political position: Government or Opposition

DATASETS

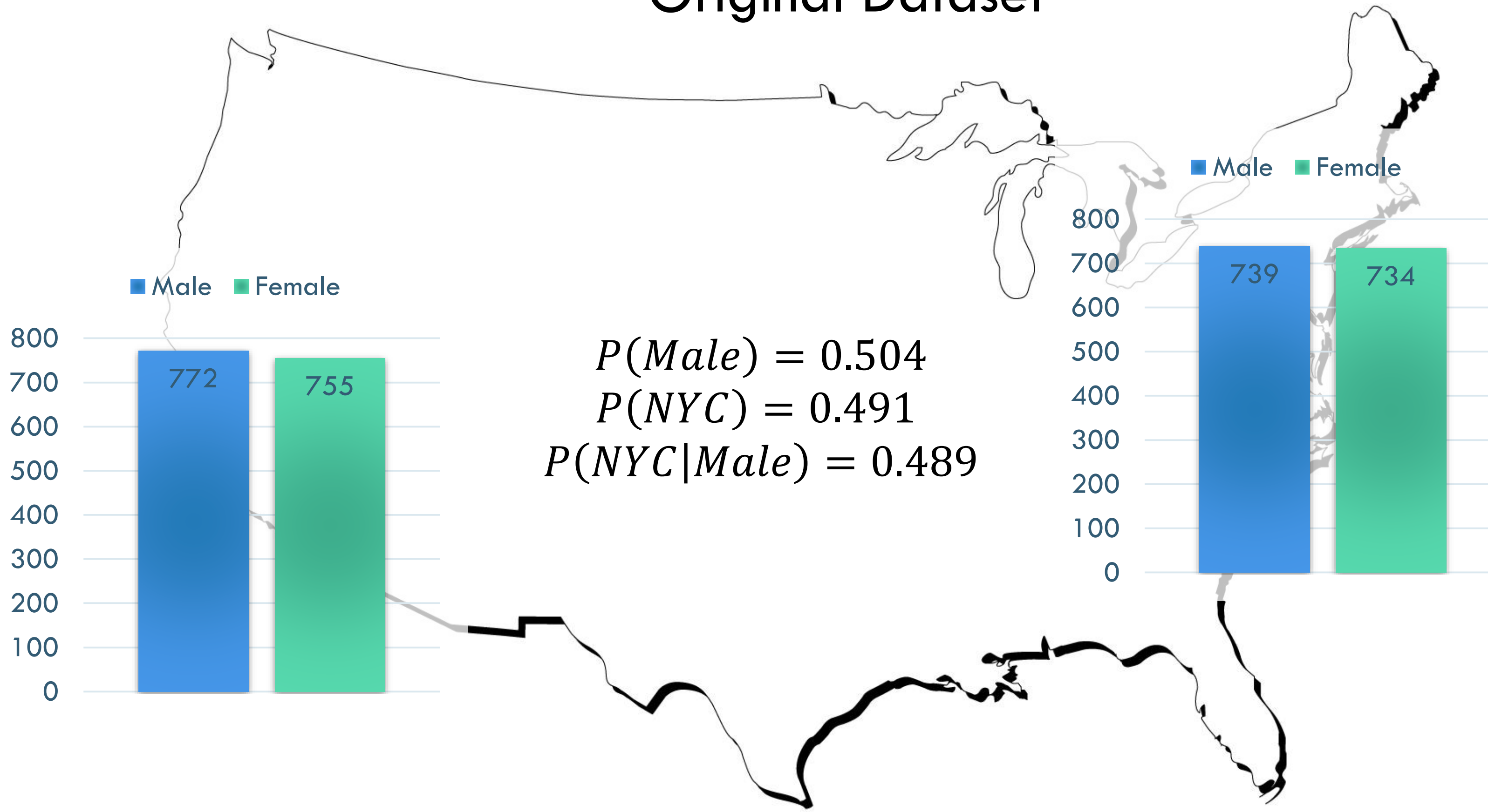
3 different datasets to experiment with back-door adjustment.

INJECTING CONFOUNDING BIAS

- Introduce confounding bias according to the following constraints:
  - $P_{train}(Y) = P_{test}(Y)$
  - $P_{train}(Z) = P_{test}(Z)$
  - $P(Y = 1|Z = 1) = b$

INJECTING CONFOUNDING BIAS

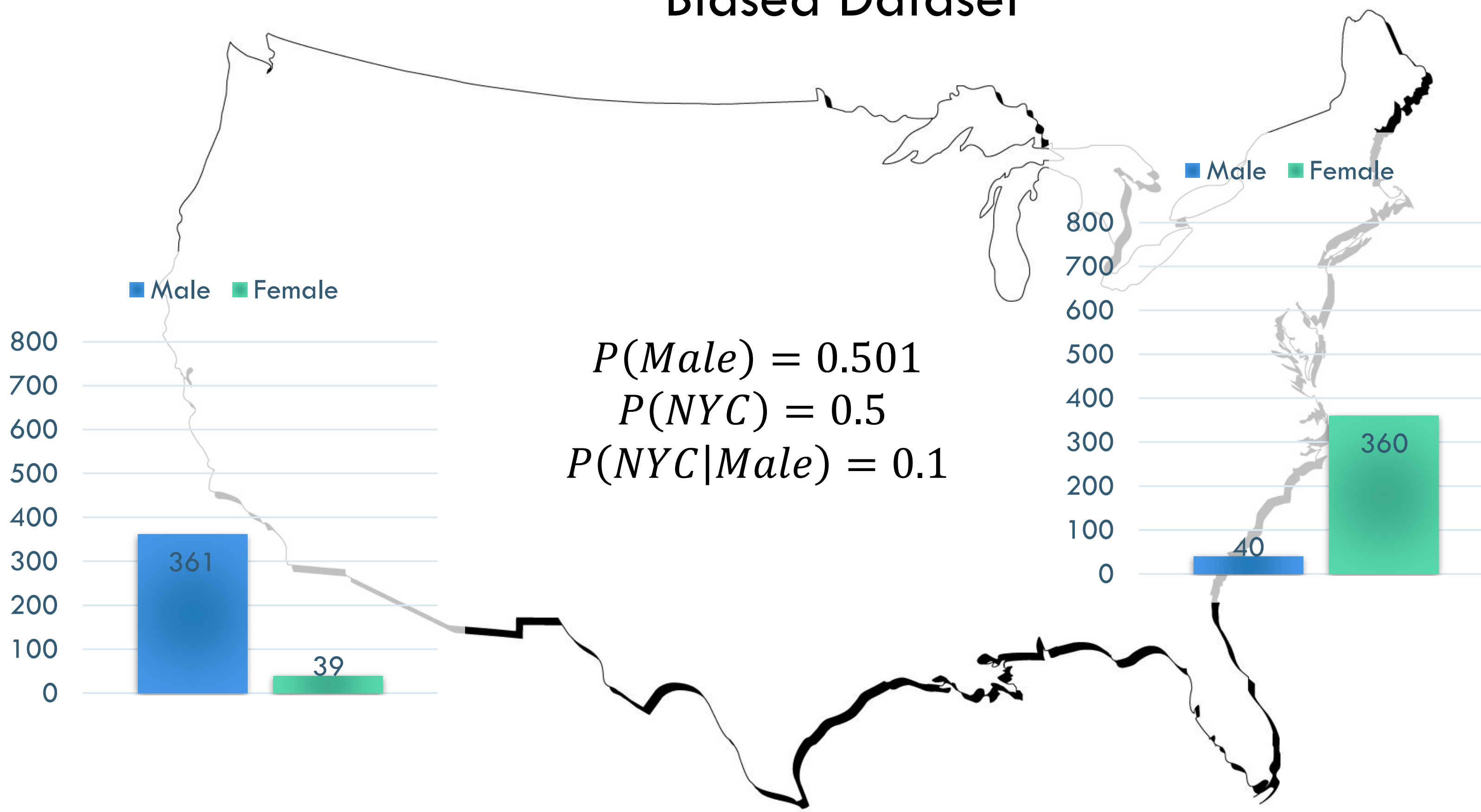
Original Dataset



$$P(Male) = 0.504$$
$$P(NYC) = 0.491$$
$$P(NYC|Male) = 0.489$$

INJECTING CONFOUNDING BIAS

Biased Dataset



$$P(Male) = 0.501$$
$$P(NYC) = 0.5$$
$$P(NYC|Male) = 0.1$$

BASELINES

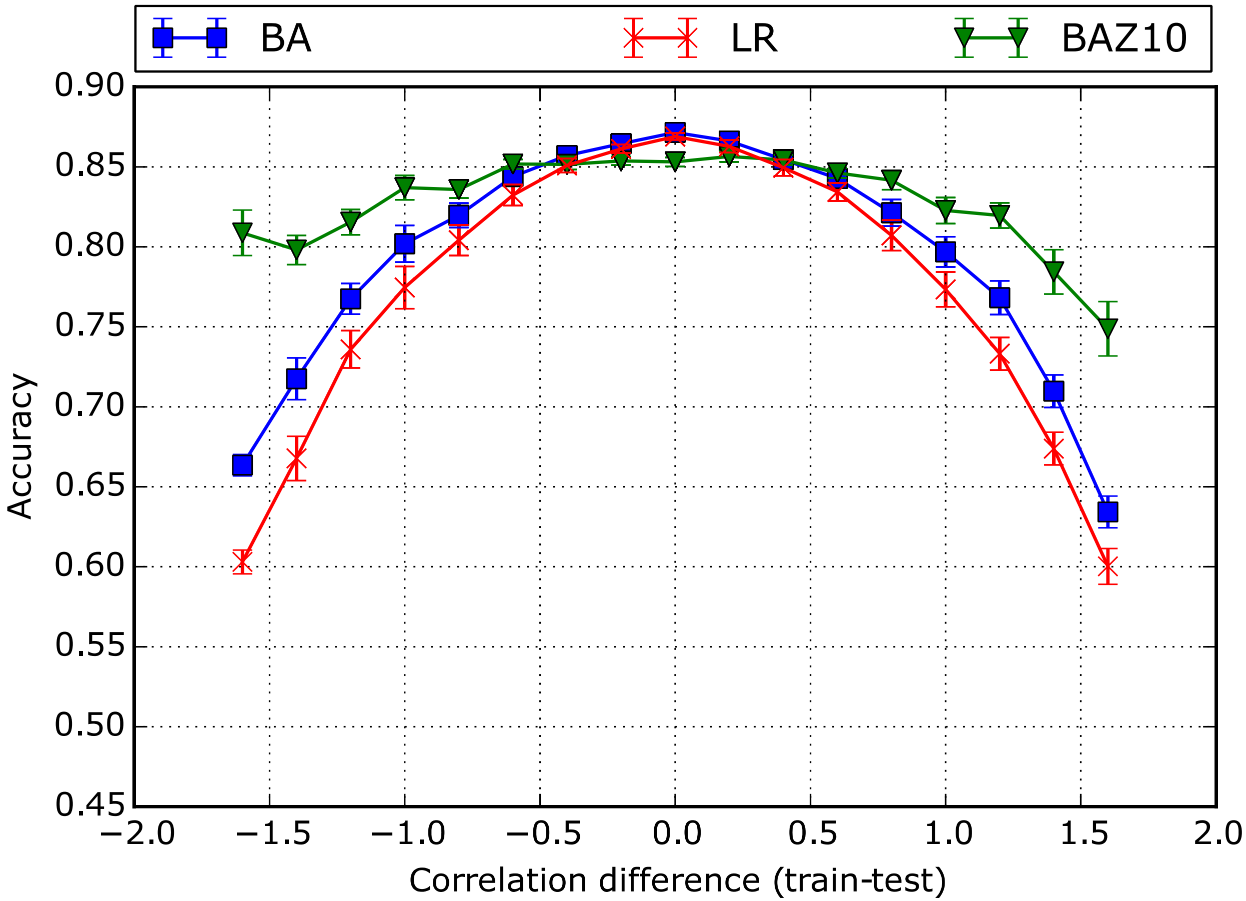
- Back-door Adjustment (BA and BAZ10)

$$L(D,\theta) = \sum_{i \in D} \log p_\theta(y_i | \boldsymbol{x}_i, z_i) - \lambda_x \sum_k (\theta_k^x)^2 - \lambda_z \sum_k (\theta_k^z)^2$$

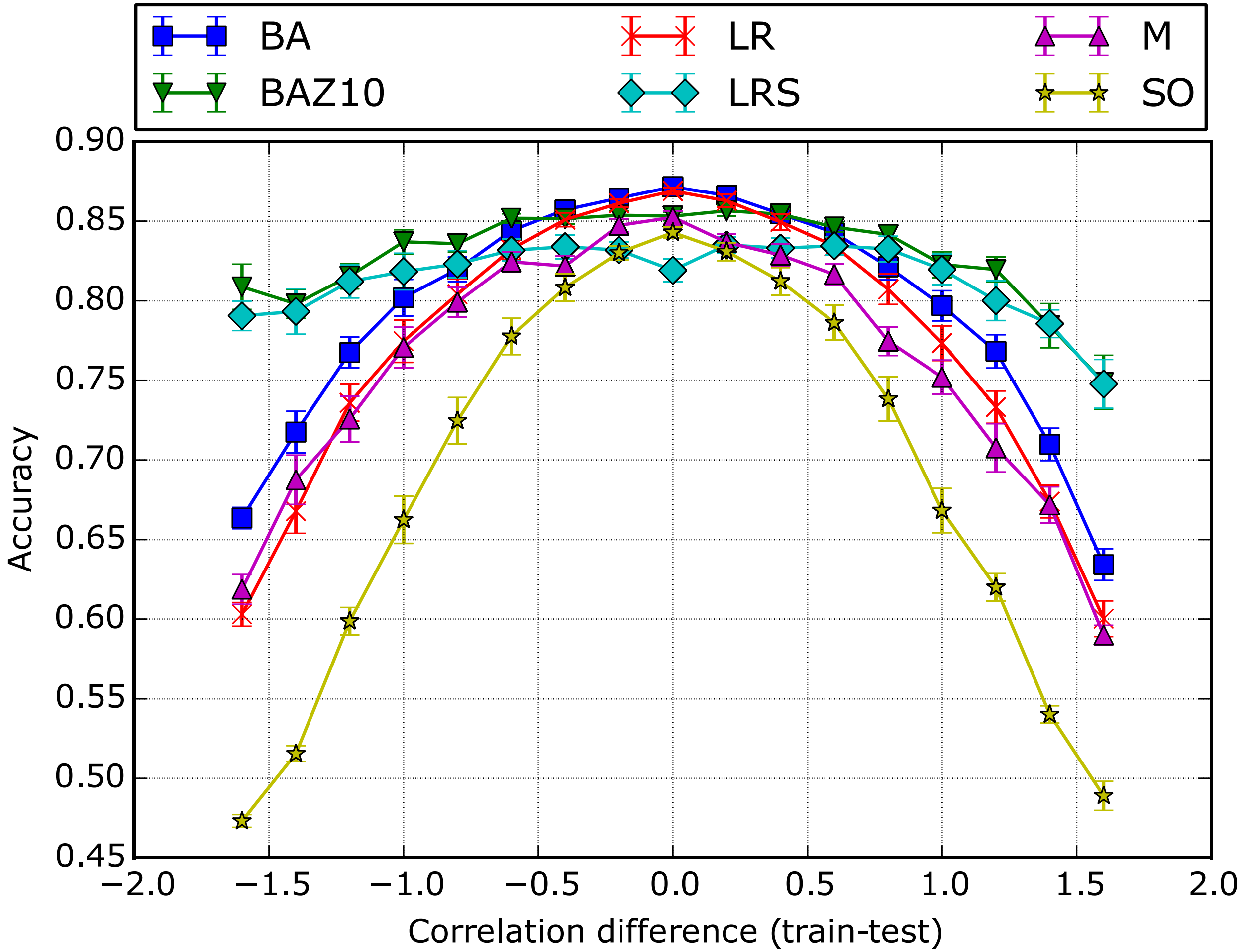
with $\lambda_z < \lambda_x$.

- Logistic Regression (LR)
- Matching (M)
- Subsampling (S)
- Sum Out (S)

RESULTS FOR THE TWITTER DATASET

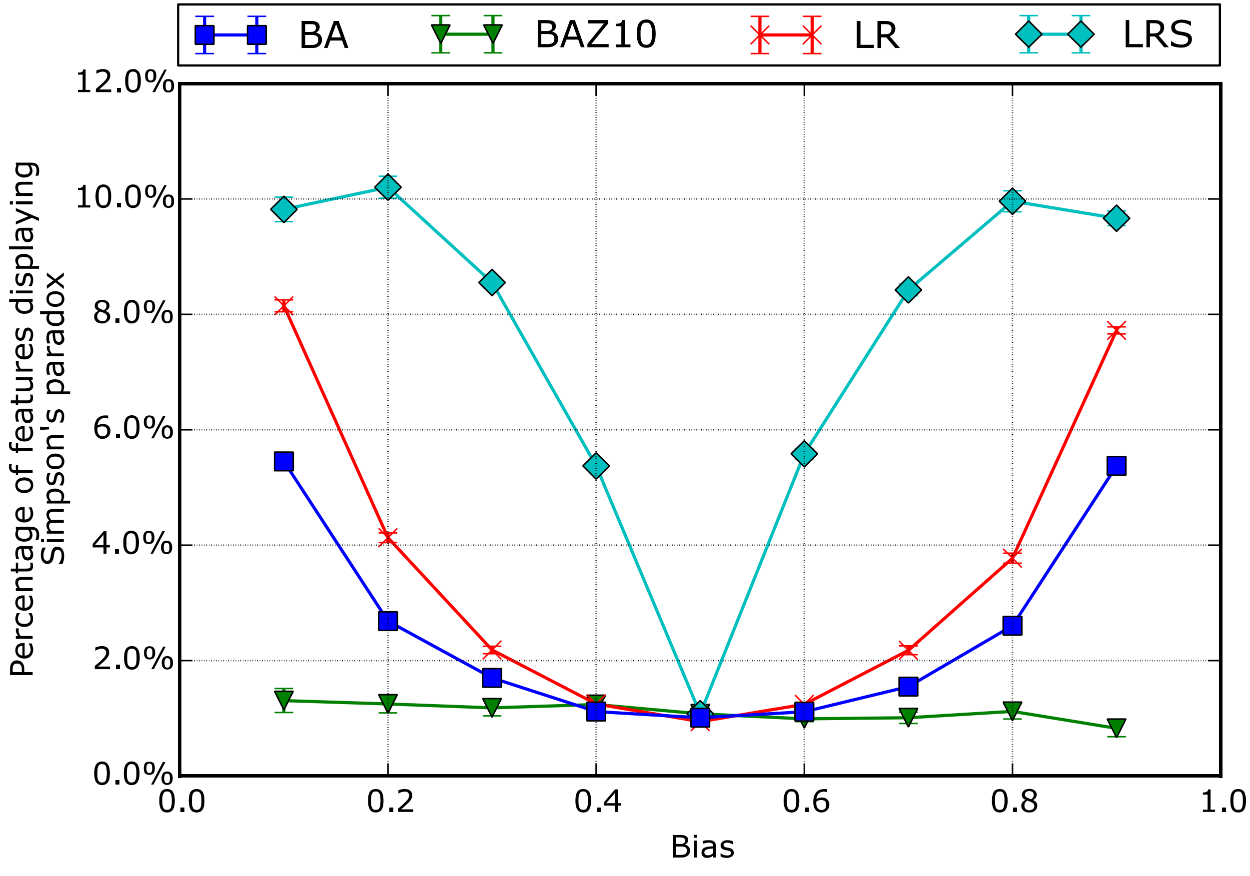


RESULTS FOR THE TWITTER DATASET



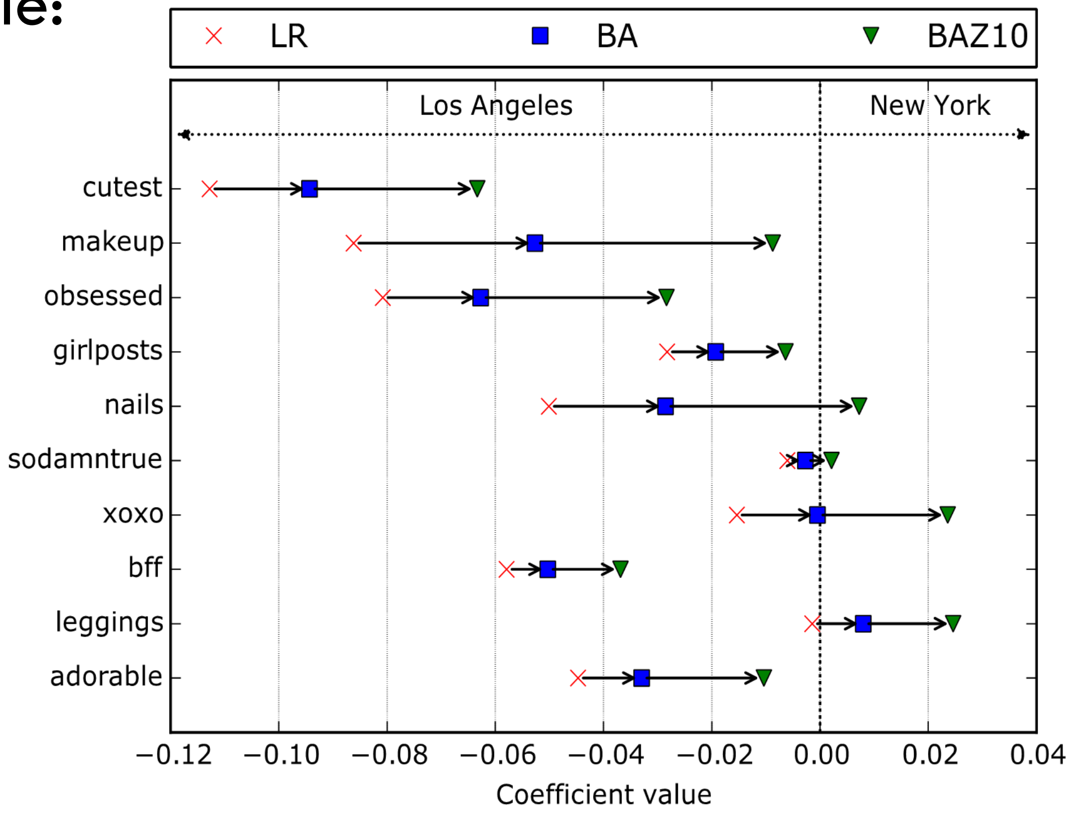
EFFECTS OF BACK-DOOR ADJUSTMENT

- Simpson's Paradox



EFFECTS OF BACK-DOOR ADJUSTMENT

- Coefficients of features predictive of the confounding variable:



CONCLUSION / FUTURE WORK

- Efficient and effective method to use back-door adjustment in text classification.
- Use back-door adjustment with a vector of confounders.
- Use back-door adjustment with a noisy measurement of the confounder.