Robust Cyberbullying Detection with Causal Interpretation

Lu Cheng, Ruocheng Guo, Huan Liu

Arizona State University

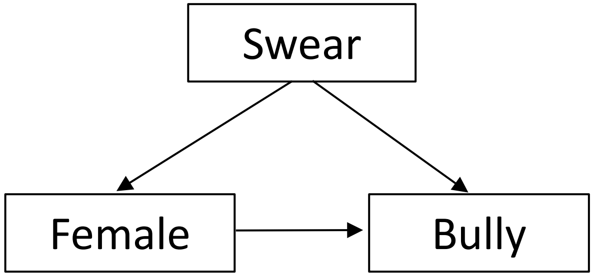
http://www.public.asu.edu/~lcheng35/

Cyberbullying in Social Media

- Definition
  - Cyberbullying is commonly defined as the electronic transmission of insulting or embarrassing comments, photos, or videos.
- Studies show that over half of adolescents and teens are faced with cyberbullying
  - Canada: ~20%, US: ~43%, Mainland China: ~57%

Limitations of Existing Work

- High cost to label data
  - Time-consuming, labor-intensive, privacy issue
- Observational data
  - *Confounders*

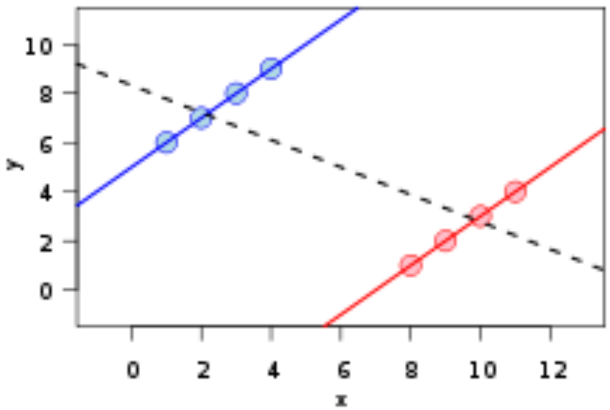


- Interpretability
  - *Causation* ≠ *Correlation*

Proposed Solution

- Simpson's Paradox
  – A phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

	Recommend Sophia's	Recommend Carlo's
Male	$\frac{50}{150} = \mathbf{30\%}$	$\frac{180}{360} = \mathbf{50\%}$
Female	$\frac{200}{250} = \mathbf{80\%}$	$\frac{36}{40} = \mathbf{90\%}$
Combined	$\frac{250}{400} = \mathbf{62.5\%}$	$\frac{216}{400} = \mathbf{54\%}$



Proposed Solution

- Simpson's Paradox
  - A compelling demonstration of the existence of confounders
  - Check Simpson's Paradox between each pair of features.

Disaggregate level

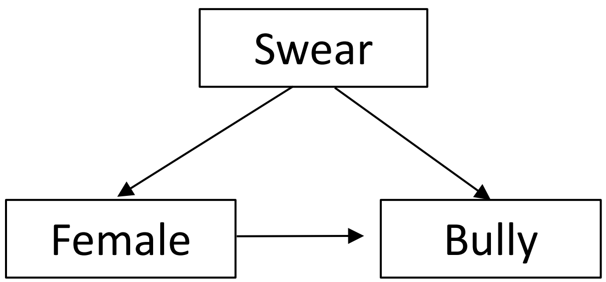
$$\mathbb{E}_c[Y|X^s, X^m] = f(\gamma + \beta_{cg} x^s).$$

Aggregate level

$$\mathbb{E}[Y|X^s] = f(\gamma + \beta_1 x^s).$$

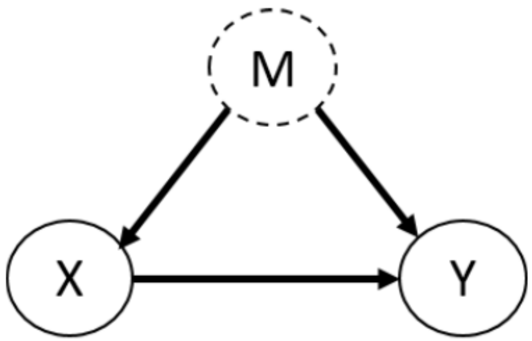
$$\Downarrow$$

$$\frac{d}{dX^s}\mathbb{E}[Y|X^s] \times \frac{d}{dX^s}\mathbb{E}_c[Y|X^s, X^m = x^m] < 0 \qquad \forall x^m.$$

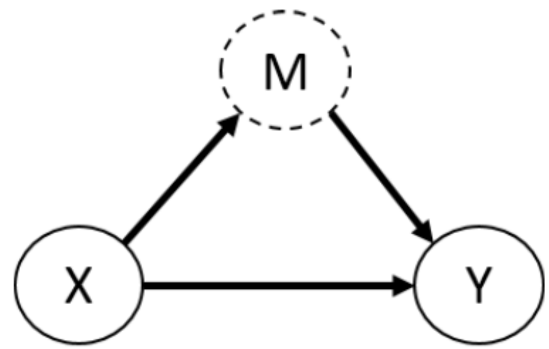


p Confounders

- The potential confounders
  - Variables that lead to most Simpson's paradox



Confounders



Causes

$$\frac{d}{dX^s}\mathbb{E}[Y|X^s]\times \frac{d}{dX^s}\mathbb{E}_c[Y|X^s, X^m = x^m] < 0 \quad \forall x^m.$$

p confounder

Simpson's pair

$$(X^s, X^m)$$

p Confounders Cont.

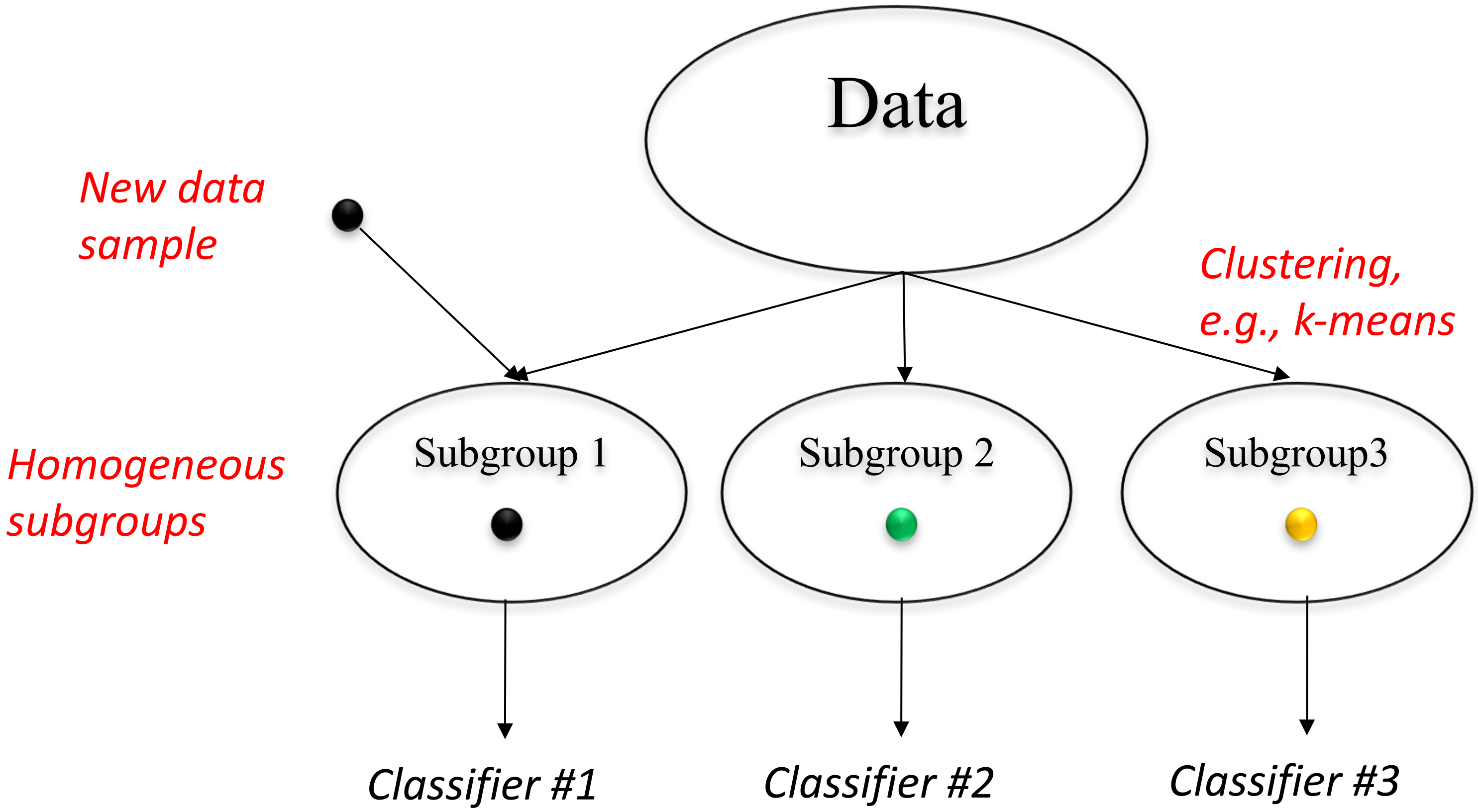
- The potential confounders
  - Variables that lead to most Simpson's paradox
  - Given all the Simpson’s pairs, count the how many times a variable is in the Simpson’s pairs
  - Set a threshold τ, *p* confounders: $Z^m > \tau$

$$z_{mp} = \begin{cases} 1, & \text{if } X_m \text{ is in Simpson's pair } p, \\ 0, & \text{otherwise.} \end{cases}$$

$$Z^m = \sum_{p=1}^{|\mathcal{P}|} z_{mp}, \qquad \forall X^m \in \mathcal{X}.$$

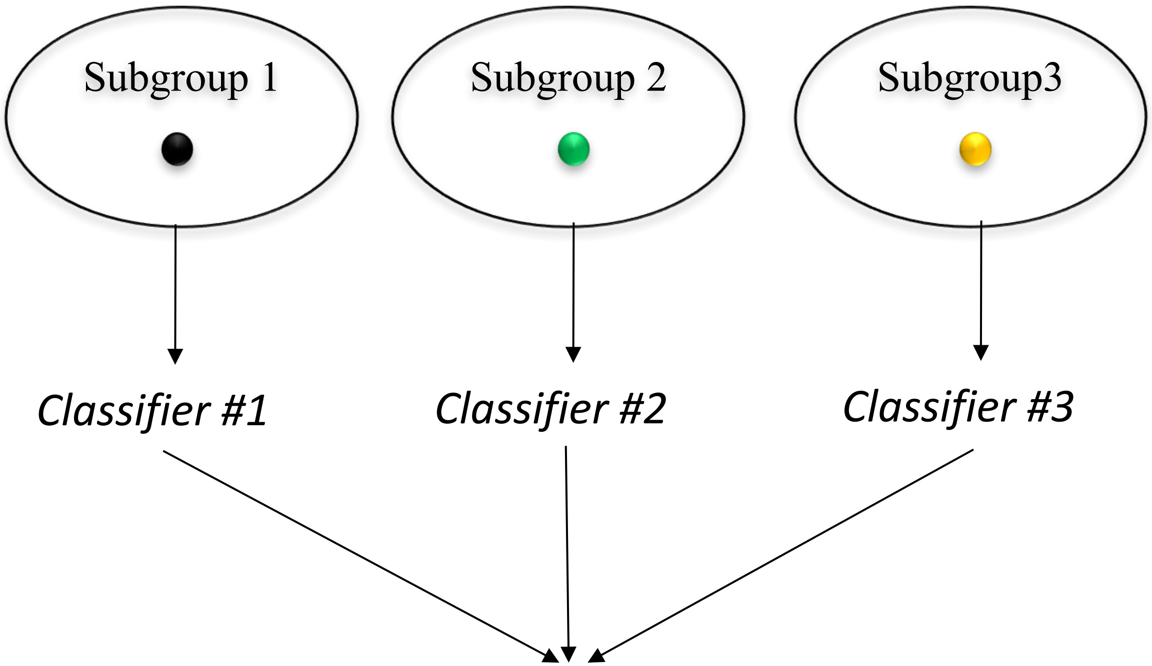
Data Disaggregation

- Clustering based on *p* confounders



Identify Potential Causes

Homogeneous subgroups



Feature Selection:
The top important features are
causal features

Evaluation Method

- Transportability/Generalizability
  - Causal relationships can be transported between different domains
  - Causal features are more robust and can generalize to changes in the data distribution
- Cross-domain cyberbullying detection
  - e.g., Dataset1 -> Dataset2

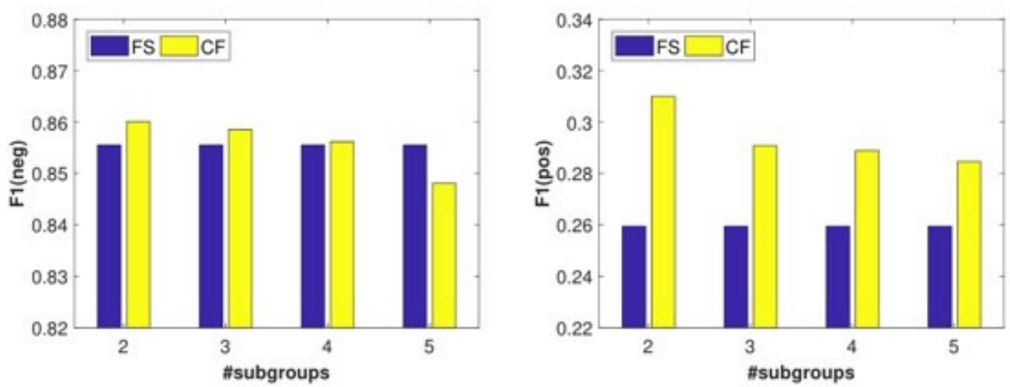
Results

- Datasets
  - Formspring
  - Twitter

Table 1: Dataset Statistic

Dataset	#Users	#Normal	#Bully	#Total
Formspring	50	12,036	1,126	13,162
Twitter	9,833	16,149	3,845	19,994

- Analysis of #subgroups



(a) F1 score (neg) (b) F1 score (pos)

Results Cont.

- Analysis of #*p* confounders

Classifiers		Random Forest				Extra Tree				AdaBoost			
#confounders		1	2	4	8	1	2	4	8	1	2	4	8
F1 (neg)	FS	0.856	0.856	0.856	0.856	0.860	0.860	0.860	0.860	**0.858**	**0.858**	**0.858**	**0.858**
	CF	**0.858**	**0.862**	**0.864**	**0.857**	**0.862**	**0.861**	**0.862**	**0.861**	0.847	0.852	0.846	0.854
F1 (pos)	FS	0.271	0.271	0.271	0.271	0.262	0.262	0.262	0.262	0.334	0.334	0.334	0.334
	CF	**0.280**	**0.311**	**0.335**	**0.315**	**0.270**	**0.286**	**0.297**	**0.297**	**0.359**	**0.371**	**0.372**	**0.346**

- Top features

Table 6: Top five important covariates of models in Group VS. Subgroups: LIWC categories Informal, Drives, Affective processes and Biological processes.

Group	Subgroup1	Subgroup2	Subgroup3
swear	power	sexual	anx
anger	achieve	bio	drives
informal	drives	negemo	affect
negemo	health	swear	affiliation

Conclusions

- We study a novel problem of interpreting cyberbullying classifier
- We propose a model that removes potential confounding bias in cyberbullying detection
- The model performs better and can identify potential causes

Q&A