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 \(\neq\) 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.

<table>
<thead>
<tr>
<th></th>
<th>Recommend Sophia’s</th>
<th>Recommend Carlo’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>( \frac{50}{150} = 30% )</td>
<td>( \frac{180}{360} = 50% )</td>
</tr>
<tr>
<td>Female</td>
<td>( \frac{200}{250} = 80% )</td>
<td>( \frac{36}{40} = 90% )</td>
</tr>
<tr>
<td>Combined</td>
<td>( \frac{250}{400} = 62.5% )</td>
<td>( \frac{216}{400} = 54% )</td>
</tr>
</tbody>
</table>
Proposed Solution

• Simpson’s Paradox
  – A compelling demonstration of the existence of confounders
  – Check Simpson’s Paradox between each pair of features.

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

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

Disaggregate level

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

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

Confounders

\[
\frac{d}{dX^s} \mathbb{E}[Y|X^s] \neq p \text{ confounder}
\]

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

Causes

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 $\tau$, $p$ confounders: $Z^m > \tau$

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

$$Z^m = \sum_{p=1}^{\mid\mathcal{P}\mid} z_{mp}, \quad \forall X^m \in X.$$
Data Disaggregation

• Clustering based on $p$ confounders

New data sample

Homogeneous subgroups

Data

Clustering, e.g., $k$-means

Subgroup 1

Classifier #1

Subgroup 2

Classifier #2

Subgroup 3

Classifier #3
Identify Potential Causes

Homogeneous subgroups

Subgroup 1

Classifier #1

Subgroup 2

Classifier #2

Subgroup 3

Classifier #3

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

• Analysis of #subgroups

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Users</th>
<th>#Normal</th>
<th>#Bully</th>
<th>#Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formspring</td>
<td>50</td>
<td>12,036</td>
<td>1,126</td>
<td>13,162</td>
</tr>
<tr>
<td>Twitter</td>
<td>9,833</td>
<td>16,149</td>
<td>3,845</td>
<td>19,994</td>
</tr>
</tbody>
</table>

(a) F1 score (neg)  
(b) F1 score (pos)
Results Cont.

• Analysis of $p$ confounders

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

• Top features

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

<table>
<thead>
<tr>
<th>Group</th>
<th>Subgroup1</th>
<th>Subgroup2</th>
<th>Subgroup3</th>
</tr>
</thead>
<tbody>
<tr>
<td>swear</td>
<td>power</td>
<td>sexual</td>
<td>anx</td>
</tr>
<tr>
<td>anger</td>
<td>achieve</td>
<td>bio</td>
<td>drives</td>
</tr>
<tr>
<td>informal</td>
<td>drives</td>
<td>negemo</td>
<td>affect</td>
</tr>
<tr>
<td>negemo</td>
<td>health</td>
<td>swear</td>
<td>affiliation</td>
</tr>
</tbody>
</table>
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