Robust Cyberbullying Detection with Causal Interpretation

Lu Cheng
Computer Science and Engineering,
Arizona State University
lcheng35@asu.edu

Ruocheng Guo
Computer Science and Engineering,
Arizona State University
rguo12@asu.edu@asu.edu

Huan Liu
Computer Science and Engineering,
Arizona State University
huanliu@asu.edu

ABSTRACT

Cyberbullying poses serious threats to preteens and teenagers, therefore, understanding the incentives behind cyberbullying is critical to prevent its happening and mitigate the impact. Most existing work towards cyberbullying detection has focused on the accuracy, and overlooked causes of the outcome. Discovering the causes of cyberbullying from observational data is challenging due to the existence of *confounders*, variables that can lead to spurious causal relationships between covariates and the outcome. This work studies the problem of robust cyberbullying detection with causal interpretation and proposes a principled framework to identify and block the influence of the plausible confounders, i.e., *p*-confounders. The de-confounded model is causally interpretable and is more robust to the changes in data distribution. We test our approach using the state-of-the-art evaluation method, *causal transportability*. The experimental results corroborate the effectiveness of our proposed algorithm. The purpose of this study is to provide a computational means to understanding cyberbullying behavior from observational data. This improves our ability to predict and to facilitate effective strategies or policies to proactively mitigate the impact of cyberbullying.

KEYWORDS

Cyberbullying detection, Causality, Social media

ACM Reference Format:
Lu Cheng, Ruocheng Guo, and Huan Liu. 2019. Robust Cyberbullying Detection with Causal Interpretation. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316503

1 INTRODUCTION

Electronic communication has provided a new context for preteen and teenage youth to bully and harass their peers. The prevention of cyberbullying behavior, hence, has been a growing public concern in the US and abroad [9]. However, preventing cyberbullying and its related harms is complicated with the limited understanding of the causes. Within the data mining community, primary efforts toward automatic detection of cyberbullying have focused on building generic binary classifiers with high accuracy rather than causal interpretation. As data for cyberbullying detection on various social media platforms are hard to obtain, this requires that cyberbullying classifiers are robust to *confounding bias* and *selection bias* to ensure the validity of studies based on the binary classification.

Even though confounder is central to much of empirical social science, it has been mostly overlooked in standard machine learning models. This is presumably because *prediction* is the goal of machine learning models, instead of *causation*. Despite their satisfactory performance, these models fail to explain the *causes* behind each prediction. The strong statistical association between the input covariates (e.g., text features) and the outcome (class label) can benefit from another covariate that is correlated with both the input and the output. For example, previous studies [17] reveal the strong correlation between Facebook intensity (a measurement of Facebook usage) and cyberbullying, but does this imply that high Facebook intensity cause cyberbullying? Studies have shown that this is spurious association due to the influence of user's age and gender, i.e., confounders [17]. Understanding confounders can also encourage generalizable science [21]. For cyberbullying detection, as data is hard to collect, conclusions that are obtained in one experimental setting should be effectively applied to another similar environment, i.e., *causal transportability* [21]. Indeed, if influence of confounders is consistent from training to testing data, the predictions should not be harmed by the presence of confounders. However, this is not true in practice as on one hand, training sets are typically small due to the annotation cost, which tends to change the data distributions. On the other hand, in many domains, relationship between the confounder and the outcome is likely to shift over time, resulting in poor accuracy. Without proper control of confounding bias, we can easily reach problematic conclusions.

Building a robust cyberbullying classifier remains a challenging task, mainly because: (1) It is often impossible to identify causal structures from pure observational data, where considerable uncertainty presents in the data generating process. A common relaxation is to find the covariates that are associated with the outcome conditioning on other observed covariates [12]. (2) Online data from social media platforms is often sparse, noisy, heterogeneous and high-dimensional, leading to tremendous spurious associations among covariates. While researchers have analyzed growing social data to understand social behavior in online platforms, it presents multi-faceted challenges that current machine learning models are not well-equipped to handle [1]. To achieve the goal of robust cyberbullying detection with causal interpretation, in this paper, we study a novel problem of controlling confounding bias in cyberbullying detection based on observed social media data. This essentially enables us to identify *plausible* confounders (*p*-confounder) and covariates that are causally related to cyberbullying behavior.

The contributions of this work are highlighted below:
(1) We study the problem of robust cyberbullying detection via the identification and control of *p*-confounders. To the best of our

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316503

knowledge, this is the first work aimed at leveraging the connections between machine learning and causality to build a robust and causally interpretable cyberbullying classifier.
(2) We develop a principled statistical method to identify p-confounders in cyberbullying detection. To block influence of these confounders, we apply clustering algorithm to stratify data into more homogeneous subgroups and classify new instances accordingly.
(3) We perform extensive experiments on two real-world datasets to examine models' robustness based on *causal transportability*, the state-of-the-art evaluation method of learning confounders from the observational data.

2 PROBLEM STATEMENT

Given a corpus of N social media posts, we denote as $\mathbf{y} = \{y_1, ..., y_N\}$ the labels of the posts with $y_i \in \{0, 1\}, i = \{1, 2, ..., N\}$, where $y_i = 1$ denotes that the post is a cyberbullying message, otherwise $y_i = 0$. $\mathbf{X}$ is the covariate matrix and its i-th column, X^i denotes the i-th covariate. The problem of robust cyberbullying detection with causal interpretation is to identify and block the influence of a group of p-confounders $\mathcal{M} = \{..., X^m, ...\}, X^m \in \mathcal{X}$ in cyberbullying detection.

To this end, we aim to answer:
How to effectively detect and block the influence of plausible confounders from the observational cyberbullying data?
Following the convention in the literature [29], we first address two required assumptions:

- *Unconfoundedness assumption.* All the covariates affecting both the input covariates X and the cyberbullying outcome Y are observed.
- *No complex causes.* A cause of cyberbullying behavior is as simple as being represented as a single covariate.

These assumptions are important as they are the foundations of the general causal inference framework. It explains that we do not consider uncertainty in the observational data due to unobserved confounders. It also assumes that no cause is composed of other causes.

3 PROPOSED FRAMEWORK

In this section, we describe the details of how to find confounder candidates and identify p-confounders.

3.1 Generate confounder candidates

Cyberbullying study can be susceptible to confounding bias, where a statistical association is distorted and does not reflect the corresponding causal relation. Confounding bias can lead to Simpson's paradox, a phenomenon whereby the association coefficient between a covariate and the outcome reverses sign upon conditioning on a second covariate, regardless of the value of the second covariate [20]. Simpson's paradox is often presented as a compelling demonstration of the existence of confounders [19] as it implies that there might be multiple causal pathways leading to the outcome. For cyberbullying detection, one of the two possible causal graphs (we assume there exist causes among the input covariates) that can result in Simpson's paradox is shown in Fig. 1. Here M is a confounder because it is on the *back-door path* $X \leftarrow M \rightarrow Y$. Hence, the causal relation between X and Y is spurious as M can influence both the main covariate X and the outcome Y through $P(X|M)$ and $P(Y|M)$.

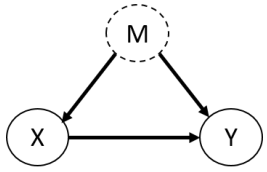


Figure 1: A causal graph leads to the Simpson's paradox where M is the confounder of the causal relationship $X \rightarrow Y$. X and Y are the main covariate and the outcome, respectively.

To detect Simpson's paradox, we first compute the statistical associations between $X^s(X)$ and Y with linear generalized linear model:

$$\mathbb{E}[Y|X^s] = f(\gamma + \beta_1 x^s), \tag{1}$$

where β_1 measures the effect of the main covariate X^s on Y, and f is a monotonically increasing function (e.g. logistic sigmoid function in this work) of its argument. Secondly, we seek for Simpson's pairs $(X^m(M), X^s)$ such that the association between X^s and Y reverses sign upon conditioning on X^m, i.e., Simpson's paradox. Specifically, conditioning on the second covariate X^m, we stratify the data to investigate the trend in each subgroup $g \in G$:

$$\mathbb{E}_c[Y|X^s, X^m] = f(\gamma + \beta_{cg} x^s). \tag{2}$$

Then the following equation should hold to ensure the reversal:

$$\begin{aligned} &\frac{d}{dX^s}\mathbb{E}[Y|X^s]\times \\ &\frac{d}{dX^s}\mathbb{E}_c[Y|X^s, X^m = x^m] < 0 \qquad \forall x^m, \end{aligned} \tag{3}$$

where the first term stands for the gradients of models at the aggregate level of data and the second term is the gradient of model in subgroups conditioning on X^m. Next, to quantify how significant these variables are correlated to each other, we perform a goodness of fit test at the significance level $\alpha = 0.05$. The test statistic of the logistic regression using the Wald test [33] is simply

$$Z = \frac{\hat{\beta}_i}{s.d.(\hat{\beta}_i)} \sim \mathcal{N}(0, 1) \quad \beta_i \in \{\beta_1, \beta_{cg}\}, \tag{4}$$

where $s.d.$ is the standard deviation. Then the obtained p-value determines "likel" or "unlikely" to reject the null hypothesis $\beta_i = 0$, i.e., if the p-value is less than (or equal to) α, then β_i is statistically different from zero and vice versa.

In summation, to observe Simpson's paradox, we need to ensure that

- Reversed sign: $\beta_1 \times \beta_{cg} < 0, \forall g \in G$, satisfying Eq. (3).
- Goodness-of-fit test: $p_{\beta_1} < \alpha, p_{\beta_{cg}} < \alpha, \forall g \in G$, where p_β is the p-value of β.

Now $\mathcal{P}$ includes all identified Simpson's pairs (X^m, X^s) in cyberbullying detection and any X^m appears in a Simpson's pair is a confounder candidate.

3.2 Detect p-confounders

As explained in [19], Simpson's paradox can result from multiple causal graphs, i.e., the presence of Simpson's paradox is a necessary

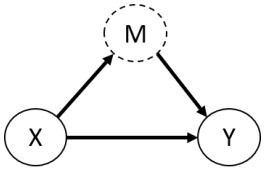


Figure 2: The other causal graph that results in Simpson's paradox in cyberbullying detection. M is not a confounder.

but not sufficient condition for the existence of confounders. In the problem setting of cyberbullying detection, two causal graphs, as shown in Fig. 1-2, can lead to Simpson's paradox as the causal directions between M and Y, X and Y can only be $X \rightarrow Y$ and $M \rightarrow Y$. Comparing the two causal graphs, M in Fig. 1 is the confounder as it influences both X and Y while M in Fig. 2 is not.

This brings up the critical question for us to find out the more likely confounders in cyberbullying detection: *Which causal graph does a Simpson's paradox may imply?* In other words, should we block influence of any variable that can lead to Simpson's paradox? The answer is "No" because there are no back-door paths requiring blockage in Fig. 2.

Let z_{mp} indicate whether the covariate X^m is a confounder candidate, that is

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

We first define the *confounder momentum* Z^m of a covariate X^m as

$$Z^m = \sum_{p=1}^{|\mathcal{P}|} z_{mp}, \qquad \forall X^m \in \mathcal{X}. \tag{5}$$

We then rank the potential confounders based on the *confounder momentum* Z^m in a descending order. Then the covariates with higher confounder momentum are more plausible confounders. This is intuitive as confounders are defined by the underlying causal structures. A confounder in a causal graph may not be a confounder in another causal graph as input covariate changes from X to X'. When blocking the influence of p-confounders, we seek for confounder candidates that have impact on larger group of input covariates in the observed data.

To identify the more likely causal graph behind a Simpson's paradox, we further propose a data-driven approach based on the theory of *causal transportability* [21]. The crucial question to address here is can we "transport"- or generalize- the cyberbullying classifiers from one population to the other? As predictions relying on causal relationships (de-confounded models) should be more robust to changes of data distribution, we compare the performance of regular cyberbullying classifiers and de-confounded cyberbullying classifiers through across-domain predictions. In the rest of this section, we describe the details of de-confounding procedure and how causal interpretation can be naturally revealed.

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

3.3 Identify potential causes

Conventional machine learning models seek features that can predict well. Nevertheless, this can constrain models' ability of generalization. An intuitive way to block the spurious statistical associations (de-confounding) is data stratification, i.e., disaggregating data into subgroups where more homogeneous instances are grouped together. Given a set of p-confounders $\mathcal{M}$, the data stratification procedure is

$$h : \mathcal{M} \rightarrow g, \ g \in \mathcal{G}, \tag{6}$$

where $\mathcal{G}$ is a set of clusters and h is a clustering function (e.g., Gaussian Mixture Clustering) that assigns samples with similar values of p-confounders to the same cluster. This step is to group instances with more homogeneous values of p-confounders to block their influence on both the covariates and the outcome. In the testing phase, given a new instance, it is first assigned to the closest cluster and then we use the classifier in that subgroup to predict the label for this instance. The de-confounded cyberbullying classifier is causally interpretable as it relies on causal relationships instead of statistical associations. At last, we obtain covariates with large coefficients. The top covariates of the de-confounded cyberbullying classifiers play central roles in causally interpreting the predicted results.

4 DATASETS

We consider two real-world datasets for empirical evaluation of the proposed method: the Formspring[1] dataset and the Twitter dataset. Formspring, as well as Twitter, has been rated as the top social media tool where cyberbullying most frequently occurs[2]. We crawled the Twitter dataset via the Twitter streaming API [15] from September 19th to 25th (2017) with the following keywords as suggested by [16]: *nerd, gay, loser, freak, emo, whale, pig, fat, wannabe, poser, whore, should, die, slept, caught, suck, slut, live, afraid, fight, pussy, cunt, kill, dick, bitch.* 20,000 tweets were manually labeled by two psychologists and a third expert was asked to resolve the conflicts. The initial agreement of the two annotators is 80%. After conflict resolution and data cleaning, we finally obtained the Twitter dataset with 19,994 labeled tweets. Table 1 shows the basic statistics of the two datasets. Code and dataset will be released upon the acceptance of the manuscript.

For the input covariates, we adopt the representation of Linguistic Inquiry and Word Count (LIWC) [22] developed for psychometric analysis. Specifically, LIWC counts words that belong to certain categories in psychology. For example, the word "cry" belongs to five categories: sadness, negative emotion, overall affect, verb and past tense verb. The results of previous research show that such

[1] Available at http://www.chatcoder.com/DataDownload

[2] http://www.foxnews.com/tech/2010/10/07/meanest-sites-prevent-cyberbullying-online-kids.html

Table 2: A sample of LIWC output variable information.

Category	Abbrev	Examples	#words
Affective processes	affect	happy,cried	1,393
Positive emotion	posemo	love,nice	620
Cognitive processes	cogproc	cause,know	797

psychometric analysis can improve the performance of cyberbullying detection [16]. For each data point, the output of LIWC is a numerical vector with covariates from 100 subcategories such as psychological processes and cognitive processes. Table 2 shows a sample of variable information output from LIWC[3].

5 RESULTS

In this section, we design experiments to compare the causal transportability of regular and de-confounded cyberbullying classifiers, and also present some qualitative analyses for further illustration. In particular, the model is trained on one dataset and tested on the other. The experiments are executed on Python 2.7 and R 3.5.1 using a Mac OS X system with Intel Core i5 and 8GB of RAM. The following experimental results are reported with 10 runs on the two datasets.

5.1 Evaluation of p-confounders

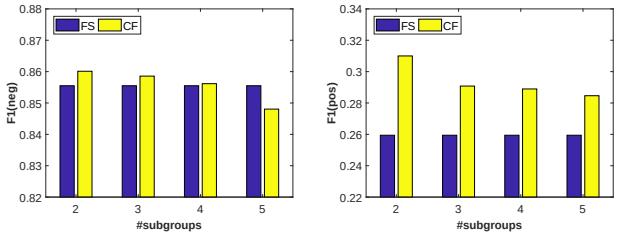
The motivation to detect p-confounders is to remove the spurious associations and identify the underlying causal mechanism of cyberbullying detection. A de-confounded model is also more robust to confounding bias and can be generalized to real-world applications [18]. As it typically takes lots of efforts to obtain real-world data for studying cyberbullying, identification of covariates that are causally related to cyberbullying can facilitate related studies on different social media platforms and digital communication tools.

To evaluate the effectiveness of the p-confounders (covariates with high confounder momentum), we compare the causal transportability of the model at the aggregate level to that at the disaggregate level of data. Specifically, we first use the K-means clustering algorithm to separate the data into K different subgroups based on the p-confounders. Then for a given testing instance, the first step is to find its most similar subgroup and then to make prediction using the classifier trained in that subgroup. The training dataset and the testing dataset should come from different domains but are sufficiently similar to each other, e.g., Twitter and Formspring.

But how to set K, the number of subgroups? The question is critical to answer as it decides how the p-confounders will be controlled and how new data instances will be classified. Here, we investigate the influence of K on the causal transportability of the de-confounded classifiers and decide the optimal K. We employ Random Forest algorithm and set K to {2,3,4,5}. We compare the results to the regular Random Forest classifier (without de-confounding) and report F1 scores for both positive (bullying) cases and negative (normal) cases due to the data imbalance. Fig. 3-4 show the results of across domain cyberbullying detection.

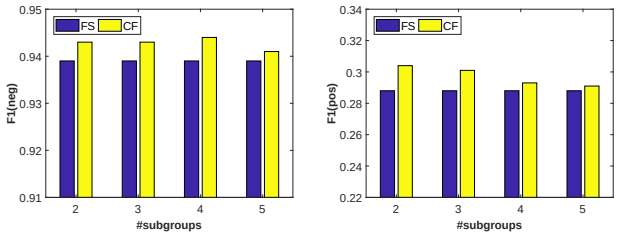
The results show that classifier with de-confounding (CF) outperform the regular classifier (FS) in most cases even when CF is

[3]Details can be seen at http://liwc.wpengine.com/



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

Figure 3: Analysis of #subgroups: Formspring → Twitter



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

Figure 4: Analysis of #subgroups: Twitter → Formspring

trained on less data due to data disaggregation. The advantages are gradually narrowed down as K increases as data disaggregation sacrifices the number of data for training. The trade-off between the number of training instances and confounding bias control should be studied in future work. Another strength of the de-confounded Random Forest classifier is it shows significant power in the prediction of bullying cases even though the training dataset is imbalanced with much more negative instances. This indicates that data stratification conditioning on p-confounders effectively mitigate influence of confounders and thus the de-confounded model is more robust to the domain changes.

With K set to 2, another research question to be addressed is *how many p-confounders are there?* Here we propose a data-driven approach. Given a ranked list of potential confounders, we use top t confounder candidates to disaggregate data. We then set $t = \{1, 2, 4, 8\}$ and test the causal transportability with three common classification models: Extra Trees classifier that fits a number of randomized decision trees (a.k.a. extra-trees) [14], Random Forest [3] and AdaBoosting [25]. Results are shown in Table 3-4, where CF denotes as the de-confounded classifiers. We can observe from the results that

- In most of the cases, the causal transportability of classifiers trained on each strata of the disaggregated data outperform regular machine learning models. For example, the de-confounded Random Forest algorithm in Table 3 outperforms the regular Random Forest by up to 23.6% in predicting bullying cases. This reveals that (1) covariates with high confounder momentum are p-confounders and (2) the proposed algorithm can effectively block influence of detected p-confounders as well.

Table 3: Robustness comparisons w.r.t. #confounders: Formspring → Twitter

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**

Table 4: Robustness comparisons w.r.t. #confounders: Twitter → Formspring

Classifiers		Random Forest				Extra Tree				AdaBoost			
#confounders		1	2	4	8	1	2	4	8	1	2	4	8
F1 (neg)	FS	0.942	0.942	0.942	0.942	0.938	0.938	0.938	0.938	0.918	0.918	0.918	0.918
	CF	**0.943**	**0.945**	**0.945**	**0.944**	**0.944**	**0.942**	**0.941**	**0.942**	**0.934**	**0.935**	**0.934**	**0.930**
F1 (pos)	FS	0.298	0.298	0.298	0.298	0.275	0.275	0.275	0.275	0.311	0.311	0.311	0.311
	CF	**0.308**	**0.309**	**0.309**	**0.306**	**0.291**	**0.292**	**0.326**	**0.299**	**0.326**	**0.328**	**0.333**	**0.332**

Table 5: Robustness comparisons: Covariates with highest confounder momentum VS. lowest ones

#confounders		1	2	4	8
Formspring→Twitter	CF	**0.270**	**0.286**	**0.297**	**0.297**
	LC	0.265	0.265	0.267	0.262
Twitter→Formspring	CF	**0.291**	**0.292**	**0.326**	**0.299**
	LC	0.279	0.277	0.285	0.280

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

- While the gain from de-confounding on predicting the negative class is not obvious, the performance of cross-domain cyberbullying detection for the positive class is improved significantly. The cyberbullying datasets in our experiments, as well as in the real world, are typically imbalanced. With the number of negative (normal) instances much larger than that of the positive (bully) instances, it is more challenging to train a classifier with high true positive than that with high true negative. However, our experiments indicate that by controlling the confounding bias, the model can find the covariates that are causally related to cyberbullying behavior. Hence it boosts the model's performance of effectively identifying cyberbullying cases.
- As the number of p-confounders used for data disaggregation increases, the de-confounded model is more capable of predicting positive cyberbullying instances. The optimal t is between 4 and 8, hence the number of p-confounders is set to $t = [4, 8)$. We conclude that covariates with large confounder momentum are effective p-confounders and blocking the influence of these covariates can improve the performance of cyberbullying detection.

To further test the effectiveness of our algorithm, we randomly pick {1,2,4,8} covariates with lowest confounder momentum (LC) and repeat the experiments in Table 3-4 using the Extra Tree classifier. In Table 5, we only report the F1 score for positive cases due to space limitation.

The results show that the Extra Tree classifier trained on the data conditioning on covariates with higher confounder momentum consistently outperform that conditioning on covariates with lower confounder momentum. This provides us additional supporting evidence for the effectiveness of the identified confounders in cyberbullying detection.

5.2 Qualitative study

Top covariates Top covariates are important because they take main responsibility for predicting cyberbullying behavior. Top covariates in a de-confounded model can causally interpret cyberbullying behavior and enhance the trust in a predictive model [27]. Here we compare the top covariates identified by Extra Tree algorithm with/without de-confounding. In this experiment, we set $K = 3$. The results are listed in Table 6.

We highlight words in different categories with different colors. We can observe that model without de-confounding (Group) identifies online *Informal* words as the most important covariates while de-confounded models detect *Drives, Affective processes, Biological processes* as the top important covariates. Interestingly, studies in psychology have revealed that cyberbullying is an imbalance of power, an urgent, basic, or instinctual need, i.e., biological drives [28]. [32] also shows that overweight and obesity, i.e., biological processes, are the key factors for school-aged children to become victims of bullying. These findings from interdisciplinary studies further corroborate that the proposed de-confounding algorithm is

Table 7: Ten Simpson's pairs in each dataset

Formspring	Twitter
(Tone, differ)	(affect, reward)
(anx, female)	(sad, achieve)
(posemo, differ)	(cogproc, female)
(swear, netspeak)	(bio, feel)
(home, negemo)	(body, sexual)
(negemo, affect)	(health, female)
(insight, Tone)	(health, body)
(tentat, affect)	(drives, Clout)
(death, negemo)	(informal, female)
(sexual, risk)	(swear, female)

effective at identifying plausible causes of cyberbullying behavior. Our algorithm shows the potential to facilitate interdisciplinary collaborations to address the societal challenge of cyberbullying.

Simpson's pairs We also provide some qualitative analyses to gain a deeper understanding of Simpson's paradox in cyberbullying detection. Table 7 includes five Simpson's pairs (potential confounder, main covariate) in both datasets. Here, *affect* represents the affective processes, *differ* represents differentiation such as *but, else*, *anx* is the abbreviation of anxiety, *cogproc* denotes cognitive process, *bio* is the abbreviation of biological process, *swear* includes swear words such as fuck, shit.

Interestingly, Table 7 reveals some counterintuitive findings. For example, studies in public health has concluded that cyberbullying is common among college women [30], words related to female should be a positive predictor of cyberbullying as such. However, our study shows that this conclusion might not be true without controlling for confounders. Specifically, for the Simpson's pair (swear, female), the trend between the *female* and cyberbullying is positive at the aggregate level while it becomes negative when conditioning on the number of *swear* words in the text. Another counterintuitive example is the pair (affect, reward). Intuitively, the feature *reward* is negatively related to cyberbullying while experimental results reveal that the trend reverses when social media users going through different *affective* processes,

6 RELATED WORK

Cyberbullying Detection Existing work of cyberbullying within computer science have primarily focused on building high-accuracy cyberbullying classifiers based on feature engineering. For example, there is a large body of work on developing textual features for detecting cyberbullying behavior [4, 11, 23, 24, 26, 31, 36]. Some work measures the content as the number of offensive terms and studies the changes of words and acronyms used in cyberbullying [24, 31]. Dinakar et al. [8] concatenated TF-IDF features, POS tags of frequent bigrams, and profane words to detect cyberbullying behaviors. Xu et al. [34] presented several off-the-shelf tools such as Bag-of-Words models and LSA- and LDA-based representation learning to predict bullying traces in Twitter. Dani et al. [7] proposed the model to incorporate sentiment information into content features. Their goal was to facilitate cyberbullying detection by capturing the sentiment consistency of normal and bullying posts. In [36], the authors took advantages of visual cues such as features extracted from images and videos to augment the accuracy of cyberbullying detection. Cheng et al. [6] leveraged the multi-modal context of social media platforms and learnt embeddings for social media sessions using heterogeneous network embedding models. In [5], the authors seek to examine cyberbullying detection based on temporal analysis of a corpus of Instagram sessions. Despite its importance, little work in computer science studied the causal interpretation and robustness of a cyberbullying classifier. A similar problem is discussed in [35]. The authors formulate the problem as a sequential hypothesis testing problem and add text-based features based on the feature scores.

Interpretable Machine Learning Our work is also related to interpretable machine learning. The volume of research in interpretability has been quickly growing along with the dominance of Deep Neural Networks. A straightforward approach is to use interpretable models such as linear regression, logistic regression and decision trees to fit data. Then features with large coefficients play key roles in interpreting the predictions. However, the forms of real-world data are usually too complicated to be modeled with simple model class. To explain the predictions of a complex machine learning model, LIME [27] tries to fit local, interpretable models that can explain single predictions of any black-box machine learning model. Koh and Liang [13] attempt to identify training points most related to a given prediction. Specifically, they use influence functions to trace a model's prediction through the learning algorithm and back to its training data. Both methods belong to post-hoc interpretability. Similar work in ad-hoc interpretability studies the feature's importance by permuting the feature's value. Based on the idea introduced in Random Forests [3], Fisher et.al [10] proposed to split the dataset in half and exchange the feature values of the two halves instead of permuting the original features. The knockoff procedure proposed in [2] aims to find truly correlated features based on the "knockoff" variables that are not associated with response but can mirror the structure of the original features.

In contrast to previous work, in this paper, we seek for covariates that are causally related to cyberbullying via blocking the influence of detected p-confounders. This is a critical issue to address because on one hand, a robust cyberbullying classifier can alleviate the problem of lacking in training data, help reach more generalizable conclusions, and on the other hand, understanding the causes of cyberbullying behavior helps improve our ability to predict and to facilitate effective strategies to proactively mitigate its impact.

7 CONCLUSION AND FUTURE WORK

We study a novel problem of robust cyberbullying detection with causal interpretation, and propose an efficient and effective algorithm to identify and block p-confounders from observational data. We first detect potential confounders using Simpson's paradox, and then identify the *most likely* confounders via a data-driven approach. Experimental results show that by controlling confounding bias, the de-confoudnded classifiers can more accurately detect cyberbullying behavior and identify covariates that are causally related to the outcome.

Our work opens several future directions. Firstly, approaches to block the influence of confounders are worth further investigation

as the current data stratification strategy sacrifices the size of training dataset. Secondly, our work assumes all the confounders are observed from data. Therefore, it may fail to consider the uncertainty from the unobserved confounders. Finally, future work should be aimed at better integrating interdisciplinary empirical findings– such as those from psychology and related social sciences–into computational models that detect cyberbullying. Interdisciplinary synergies hold particular promise for identifying, addressing, and preventing this major social problem.

ACKNOWLEDGEMENT

We thank our colleagues from the School of Mathematical and Natural Sciences and School of Social and Behavioral Sciences at ASU, who provided insight and expertise that greatly assisted the research. We would also like to show our gratitude to the 3 "anonymous" reviewers for their so-called insights.

REFERENCES

[1] Nazanin Alipourfard, Peter G Fennell, and Kristina Lerman. 2018. Can you Trust the Trend?: Discovering Simpson's Paradoxes in Social Data. In *WSDM*. ACM, 19–27.
[2] Rina Foygel Barber and Emmanuel J Candes. 2016. A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv:1602.03574* (2016).
[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[4] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*. ACM, 13–22.
[5] Lu Cheng, Ruocheng Guo, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network. In *SDM*.
[6] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. XBully: Cyberbullying Detection within a Multi-Modal Context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 339–347.
[7] Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment Informed Cyberbullying Detection in Social Media. In *ECML PKDD*. Springer, 52–67.
[8] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11, 02 (2011).
[9] L Christian Elledge, Anne Williford, Aaron J Boulton, Kathryn J DePaolis, Todd D Little, and Christina Salmivalli. 2013. Individual and contextual predictors of cyberbullying: The influence of children's provictim attitudes and teachers' ability to intervene. *Journal of youth and adolescence* 42, 5 (2013), 698–710.
[10] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2018. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the" Rashomon" Perspective. *arXiv preprint arXiv:1801.01489* (2018).
[11] Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2018. A unified deep learning architecture for abuse detection. *arXiv preprint arXiv:1802.00385* (2018).
[12] Jaime Roquero Gimenez, Amirata Ghorbani, and James Zou. 2018. Knockoffs for the mass: new feature importance statistics with false discovery guarantees. *arXiv preprint arXiv:1807.06214* (2018).
[13] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730* (2017).
[14] Manish Mehta, Rakesh Agrawal, and Jorma Rissanen. 1996. SLIQ: A fast scalable classifier for data mining. In *EDBT*. Springer, 18–32.
[15] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.
[16] Parma Nand, Rivindu Perera, and Abhijeet Kasture. 2016. " How Bullying is this Message?": A Psychometric Thermometer for Bullying.. In *COLING*. 695–706.
[17] Sara Pabian, Charlotte JS De Backer, and Heidi Vandebosch. 2015. Dark Triad personality traits and adolescent cyber-aggression. *Personality and Individual Differences* 75 (2015), 41–46.
[18] Michael J Paul. 2017. Feature Selection as Causal Inference: Experiments with Text Classification. In *CoNLL*. 163–172.
[19] Judea Pearl. 2000. *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.
[20] Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
[21] Judea Pearl and Elias Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *ICDMW*. IEEE, 540–547.
[22] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.
[23] Jing Qian, Mai ElSherief, Elizabeth M Belding, and William Yang Wang. 2018. Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection. *arXiv preprint arXiv:1804.03124* (2018).
[24] Elaheh Raisi and Bert Huang. 2017. Co-trained Ensemble Models for Weakly Supervised Cyberbullying Detection. In *NIPS Workshop on Learning with Limited Labeled Data*.
[25] Gunnar Rätsch, Takashi Onoda, and K-R Müller. 2001. Soft margins for AdaBoost. *Machine learning* 42, 3 (2001), 287–320.
[26] Imam Riadi. 2017. Detection of cyberbullying on social media using data mining techniques. *International Journal of Computer Science and Information Security (IJCSIS)* 15, 3 (2017).
[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*. ACM, 1135–1144.
[28] Ken Rigby. 2002. *New perspectives on bullying*. Jessica Kingsley Publishers.
[29] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
[30] Ellen M Selkie, Rajitha Kota, and Megan Moreno. 2016. Cyberbullying behaviors among female college students: witnessing, perpetration, and victimization. *College student journal* 50, 2 (2016), 278–287.
[31] Phoey Lee Teh, Chi-Bin Cheng, and Weng Mun Chee. 2018. Identifying and Categorising Profane Words in Hate Speech. In *ICCDA*. ACM, 65–69.
[32] Musleh uddin Kalar, Tashaba Qaiser Faizi, Maheen Jawed, Sumaira Khalil, Sara M Hussain, Mushkbar Fatima, Faiza Aslam, Hasnain Abbas Dharamshi, and Tahira Naqvi. 2015. Bullying, overweight and physical activity in school children of Karachi. *International Archives of Medicine* 8 (2015).
[33] A. Wald. 1945. Sequential Tests of Statistical Hypotheses. *Ann. Math. Statist.* 16, 2 (06 1945), 117–186. https://doi.org/10.1214/aoms/1177731118
[34] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *NAACL HLT*. Association for Computational Linguistics, 656–666.
[35] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2018. Cyberbullying Detection on Instagram with Optimal Online Feature Selection. In *ASONAM*.
[36] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network.. In *IJCAI*. 3952–3958.