2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Fairness across Network Positions in Cyberbullying Detection Algorithms

Vivek K. Singh and Connor Hofenbitzer
Rutgers University
{vs451, cfh41}@scarletmail.rutgers.edu

***Abstract*— Cyberbullying, which often has a deeply negative impact on the victim, has grown as a serious issue in online social networks. Recently, researchers have created automated machine learning algorithms to detect Cyberbullying using social and textual features. However, the very algorithms that are intended to fight off one threat (*cyberbullying*) may inadvertently be falling prey to another important threat (*bias* of the automatic detection algorithms). This is exacerbated by the fact that while the current literature on algorithmic fairness has multiple empirical results, metrics, and algorithms for countering bias across immediately observable demographic characteristics (e.g. age, race, gender), there have been no efforts at empirically quantifying the variation in algorithmic performance based on the network role or position of individuals. We audit an existing cyberbullying algorithm using Twitter data for disparity in detection performance based on the network centrality of the potential victim and then demonstrate how this disparity can be countered using an Equalized Odds post-processing technique. The results pave the way for more accurate and fair cyberbullying detection algorithms.**

Keywords- Algorithmic Fairness, Cyberbullying, Network Position

I. INTRODUCTION

In multiple domains, ranging from automatic face detection to automated decisions on parole, machine learning algorithms have been found to be systematically biased and favoring one demographic group over another [1,2,3]. This is problematic as these algorithms are amplifying existing disparities across different groups of individuals. As a result, certain groups of people may get lesser access to loans, college admissions, parole opportunities, and so on.

At the same time, the discussions around fairness (like in the scenarios above) typically rest on the notion of individual. However, much of the data being produced and the decisions being made today occur in a networked setting. As argued by *boyd* et al. [4], we must rethink our models of discrimination and our mechanisms of accountability. We need to "look beyond immutable characteristics of individuals and attend to the positions of individuals in networks" [4].

Hence, understanding the role played by one's position in a network in regard to computational algorithms is urgent and important. This work focuses on the fairness of cyberbullying detection algorithms across recipients with different network characteristics or positions. *If the algorithms works accurately when an individual with high network centrality is the potential victim and poorly when an individual with low network centrality is the potential victim, then that would be unfair.* In particular, the individuals with lower network centrality will suffer from a "double whammy" because: (1) historical research has shown that individuals on the edges of the network tend to be bullied more often than those in the center [5]; (2) those in the center of the network tend to have more data available for learning opportunities for the various machine learning algorithms. Hence, algorithms are more likely to work better for those cases where the potential victims are in the center of the network rather than those on the peripheries.

The main contributions of this work are:

(1) To motivate and ground the use of an individual's network centrality as a sensitive attribute for discrimination analysis.

(2) To audit an existing social network features based cyberbullying detection algorithm for bias based on recipient's network position and demonstrate a way to counter it.

II. RELATED WORK

Previous approaches in increasing fairness can broadly be classified into those that involve *pre-processing* the data going into the algorithms, *processing* during the prediction algorithms themselves, and those that *post-process* the results of an existing algorithm to allow for fairer decisions [1, 2, 3]. Per our knowledge, we are the first to employ fairness-based techniques while focusing on network position of a person to identify the favored and disfavored groups. The two closest related lines of works are [4] and [6]. Boyd et al., [4] argue conceptually about the roles of networks in creating biases but do not deal with empirical data. Fish et al., [6] study the problem of equal access to information as it spreads in a network but study the problem of "social welfare function", which is very different from the idea of fairness for individuals or groups when considering their specific characteristics.

The problem of cyberbullying detection has been studied in multiple domains. Dinakar et al., [7] describe cyberbullying as "when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person." Clearly, cyberbullying involves a content (text, image) component and a social component. However, most of the work on cyberbullying detection focuses on (sophisticated) textual analysis. Work by Huang et al. [8] was the first effort to identify

ASONAM '19, August 27–30, 2019, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6868-1/19/08…$15.00
https://doi.org/10.1145/3341161.3342949

the use of social features in cyberbullying detection. Since then, multiple other efforts (e.g., [9, 10]) have also used social features for cyberbullying detection. Given the importance of social aspects in cyberbullying, it is important to consider the question of fairness in terms of the network position of the recipient i.e. the potential victim.

To quantify the "fairness" of algorithms, we focus on the comparisons between privileged and underprivileged groups based on three different metrics: difference in accuracy (or AUCROC), equal odds, and equality of opportunity. Equality of opportunity (EoO) metric mandates an equal true positive rate (TPR) for the groups considered (e.g. male and female; or Low network centrality and High network centrality). Almost all practical algorithms have TPR below 100%. In such cases, EoO principle mandates a ground truth based "true" cyberbullying post should have equal odds of being labeled as "true" for cyberbullying by the detection algorithm irrespective of the network centrality of the recipient (potential victim). Equal odds metric is an extension of the above idea to include both the true positive rate and the false positive rate (FPR) [3]. Hence, the difference in the false positive rates for different considered groups (e.g. low/high centrality) is also considered in this work. The overall goal of this work is to minimize the discrepancy in the accuracy, TPR, and FPR based on the centrality of the message recipient.

III. DATASET AND APPROACH

We use a labeled cyberbullying dataset as utilized in [8]. This dataset is a subset of the Twitter corpus from the CAW 2.0 data set, which has been annotated by three labelers for the presence of cyberbullying. This data set contains 4,865 messages with 93 (roughly 2%) of them labeled as bullying messages. Like our previous work [8], the following social network features are defined for the relationship graph: (1) number of nodes, (2) number of edges, (3) degree centrality- with variants for in-degree, out-degree, sender and receiver resulting in four different features, (4) edge betweenness centrality of the edge between sender and receiver, (5) tie strength between a sender and a receiver, and (6) community embeddedness for the sender and receiver (two features), resulting in a total of ten social features to describe a user's social interactions. Similarly, based on [8] the following textual features were included: (1) density of bad words, (2) density of uppercase letters, (3) number of exclamation points and question marks, (4) number of smileys, and (5) part-of-speech-tags, these were chosen based on their correlation to the predictors output.

An important problem with cyberbullying datasets is the data imbalance. To mitigate the effects of imbalance, we applied the 'SMOTE' method [11]. To train and validate the predictions we conducted a 70%-30% split after shuffling the dataset to allow for instances of cyberbullying to be in both the training and testing set. We then applied SMOTE preprocessing on the training set. This resulted in an equal number of bullying and non-bullying instances and increased the number of instances in the training set from 3,420 to around 6,750. This allowed for more instances of the minority class to be used in training, potentially increasing the accuracy of predictions. The test set remains imbalanced to mimic the real-world scenario.

After SMOTE, we applied a dagging (Directed Aggregating) algorithm, which was the best performing algorithm in [8], to create a model for cyberbullying detection. We received confidences from the dagging predictor to be used in calculating AUROC later. We were able to obtain probability scores by using the notion of soft-voting, which is the average of the models voting rather than a hard cut off for each model.

A. Auditing algorithm for bias

We chose 'outdegree centrality for the recipient' as our sensitive attribute as this could indicate network position, which could unfairly affect a user's probability of being identified as a target for cyber bullying. As suggested in recent efforts on fair machine learning [1, 3], the sensitive attribute was not included in the algorithm's predictions as this could lead to more biases in the predictions. We calculated the median of the sensitive attribute to create two groups– those with "high" network centrality and those with "low" network centrality. Next, we audited the outputs of the algorithm for possible bias. We computed the above-mentioned algorithm's predictions, through which we were able to calculate receiver operating characteristic (AUCROC) scores as well as other performance metrics (TPR, FPR) for the two groups. Note that AUCROC is a more robust metric for measuring the performance of algorithms and is preferred to simple accuracy metric in scenarios involving imbalance across classes [11]. The above process allowed us to determine the difference in accuracy metrics across the two groups.

We ran the auditing algorithm 100 times to allow for more confidence in results and for determining statistical significance of the results. Each test round used a new random seed that was used for the test-train split, meaning that random samples were drawn from the population.

TABLE I: AVERAGE TPR, FPR, AND AUCROC COMPARISON FOR GROUPS WITH LOW AND HIGH NETWORK CENTRALITY (BASELINE).

Attribute	Baseline		
	TPR	FPR	ROC AUC
"High" network centrality	0.8102	0.3801	0.7714
"Low" network centrality	0.5328	0.1398	0.7153
Delta	0.2774	0.2403	0.0561

Throughout the analysis, we found that accuracy scores were higher when the recipients of the messages had "high" network centrality than when the recipients had "low" network centrality. We conducted a *t-test* with α=0.05 threshold for TPR, FPR and AUCROC difference between the groups. This difference was found to be ***statistically significant.***

B. Debiasing algorithm using equalized odds post-processing

Equal odds principle requires the TPR and FPR to be equal for both the underprivileged and privileged classes. Here we adapted the Equalized Odds Post-processing approach as proposed by [3] and as available in the IBM AIF 360 library [12] to compute the ROC for the considered groups. Using the AIF 360 library we implemented the classification metric class

to obtain various performance values (AUC ROC, TPR, FPR) for each group before and after the debiasing process. The library was adapted to include calculations for area under the ROC curve between two groups to better suit this paper as the original library had no notion of AUCROC.

In designing a derived predictor from binary $\tilde{Y}$ and A we can only set four parameters: the conditional probabilities $p_{ya} = Pr\{ (\tilde{Y} = 1 \mid \hat{Y} = a, A = a\}$. These four parameters, $p = (p_{00}, p_{01}, p_{10}, p_{11})$, together specify the derived predictor $\tilde{Y}_p$. For equal odds, this requires that for the outcome y, $\hat{Y}$ has equal positive rates for each group, A = 0, A = 1. Since the expected loss $El(\tilde{Y}_p, Y)$ is also linear in p, the optimal derived predictor can be obtained as a solution to the following linear program with four variables and two equality constraints:

$$El(\tilde{Y}_p, Y)$$

$$\text{s.t. } \gamma_0(\tilde{Y}_p) = \gamma_1(\tilde{Y}_p) \text{ and } \forall_{y,a} \leq p_{ya} \leq 1$$

where the components of $\gamma_a(\tilde{Y})$ are the false positive rate and the true positive rate within the considered group A = a.

IV. RESULTS

Table 2 shows the results for the chosen classification metrics after applying Equalized Odds post-processing. The comparison between the approaches (before and after the Equalized Odds post-processing) is summarized in Table 3. The results indicate that the proposed approach resulted in a lower discrepancy between the two centrality-based groups in terms of ROC AUC, TPR and FPR. These decreases in differences were validated using one-sided t-tests. The differences in scores for TPR, FPR, and AUC were found to be ***statistically significant*** at α=0.05 threshold. Note that this increase in fairness came with a slight decrease in overall AUC from 0.7434 to 0.7283, which was found to be ***not statistically significant***. Based on the trends observed we consider the proposed approach to be useful at reducing disparity in the performance of cyberbullying detection algorithms across different groups based on network centrality of the recipients.

Limitations of this work include its focus on a single cyberbullying algorithm and a single dataset. Also, a single network feature (outdegree network centrality) has been used to operationalize network position. At the same time, this work marks the first empirical effort at analyzing the difference in performance based on network position of a person – not just in cyberbullying literature but in any application domain. The results obtained here are promising and motivate further work in this direction.

TABLE II: AVERAGE TPR, FPR, AND AUCROC COMPARISON FOR GROUPS WITH LOW AND HIGH NETWORK CENTRALITY (PROPOSED APPROACH).

Attributes	Proposed Method		
	TPR	FPR	ROC AUC
"High" network centrality	0.7019	0.3339	0.7112
"Low" network centrality	0.5379	0.1427	0.7454
Delta	0.1641	0.1912	-0.0342

TABLE I: COMPARISON OF DELTAS FOR GROUPS WITH HIGH AND LOW CENTRALITY IN THE BASELINE AND PROPOSED APPROACHES.

Attributes	Deltas across high/low centrality groups		
	TPR	FPR	ROC AUC
Baseline |Delta|	0.2774	0.2403	0.0561
Proposed |Delta|	0.1641	0.1912	0.0342
Change	0.1133	0.0492	0.0119

V. CONCLUSION AND FUTURE WORK

This short paper motivates and grounds the use of network characteristics (e.g. network centrality) as a sensitive attribute to study algorithmic fairness. The audit of an existing cyberbullying detection algorithm [8] yielded that the performance of the algorithm varied quite significantly depending on the network centrality of the recipient of the potentially bullying message. This disparity in the performance was found to reduce statistically significantly with the application of the equalized odds post-processing technique. While early, the results significantly move forward the literature on fairness in networked algorithms and specifically cyberbullying detection. Future improvements on this work could consider larger network size, diverse operationalizations of network positions, and newer debiasing approaches to create fair and accurate network-centric algorithms.

ACKNOWLEDGMENT

This material is in part based upon work supported by the National Science Foundation under Grant No. IIS-1464287.

REFERENCES

[1] Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems (pp. 3992-4001).
[2] Kamishima, Toshihiro & Akaho, Shotaro & Asoh, Hideki & Sakuma, Jun. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. 35-50. 10.1007/978-3-642-33486-3_3.
[3] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323)
[4] boyd, D., Levy, K., & Marwick, A. (2014). The networked nature of algorithmic discrimination. Data and Discrimination: Collected Essays. Open Technology Institute.
[5] Festl, R., & Quandt, T. (2013). Social relations and cyberbullying: The influence of individual and structural attributes on victimization and perpetration via the internet. Human Communication Research, 39(1), 101–126
[6] Fish, B., Bashardoust, A., Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2019). Gaps in Information Access in Social Networks. arXiv preprint arXiv:1903.02047.
[7] Dinakar, K., Reichart, R., & Lieberman, H. (2011, July). Modeling the detection of textual cyberbullying. In AAAI conference on weblogs and social media.
[8] Huang, Q., Singh, V.K., & Atrey, P. (2014). Cyber bullying detection using social and textual analysis. ACM International Workshop on Socially Aware Multimedia. (pp. 3-6).
[9] Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015, August). Identification and characterization of cyberbullying dynamics in an online social network. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (pp. 280-285). ACM.
[10] Singh, V. K., Huang, Q., & Atrey, P. K. (2016, August). Cyberbullying detection using probabilistic socio-textual information fusion. In Proc. 2016 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (pp. 884-887).
[11] N. V. Chawla. (2005). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook, pages 853–867. Springer.
[12] Bellamy, R.K. et al., (2018). "AI fairness 360: An extensible toolkit for detecting understanding and mitigating unwanted algorithmic bias", *arXiv preprint arXiv:1810.01943*