Reduced-Bias Co-Trained Ensembles for Weakly Supervised Cyberbullying Detection

Elaheh Raisi* and Bert Huang

Virginia Tech, Blacksburg VA, USA
{elaheh,bhuang}@vt.edu

Abstract. Social media reflects many aspects of society, including social biases against individuals based on sensitive characteristics such as gender, race, religion, physical ability, and sexual orientation. Machine learning algorithms trained on social media data may therefore perpetuate or amplify discriminatory attitudes against various demographic groups, causing unfair decision-making. One important application for machine learning is the automatic detection of cyberbullying. Biases in this context could take the form of bullying detectors that make false detections more frequently on messages by or about certain identity groups. In this paper, we present an approach for training bullying detectors from weak supervision while reducing the degree to which learned models reflect or amplify discriminatory biases in the data. Our goal is to decrease the sensitivity of models to language describing particular social groups. An ideal, fair language-based detector should treat language describing subpopulations of particular social groups equitably. Building on a previously proposed weakly supervised learning algorithm, we penalize the model when discrimination is observed. By penalizing unfairness, we encourage the learning algorithm to avoid unfair behavior in its predictions and achieve equitable treatment for protected subpopulations. We introduce two unfairness penalty terms: one aimed at removal fairness and another at substitutional fairness. We quantitatively and qualitatively evaluate the resulting models' fairness on a synthetic benchmark and data from Twitter comparing against crowdsourced annotation.

Keywords: Cyberbullying detection · social media · weakly supervised machine learning · co-trained ensemble · fairness in machine learning · embedding models

1 Introduction

As a technology to better connect individuals, social media introduces benefits that can be nullified by the detrimental behaviors it amplifies, such as online harassment, cyberbullying, hate speech, and online trolling [2]. The serious consequences of these behaviors compels the development of automated, data-driven techniques for detecting such behaviors. A key concern in the development and adoption of machine learning for building harassment detectors is whether the learned detectors are *fair*. Most machine learning models trained on social media data can inherit or amplify biases present in training data, or they can fabricate biases that are not present in the data. Biases in harassment detectors could be characterized as when harassment detectors are more sensitive to harassment committed by or against particular groups of individuals, such

as members of ethnic, gender, sexual orientation, or age groups, which result in more false detections on protected target groups. Recent reactions to a Google Jigsaw tool for quantifying toxicity of online conversations (see e.g., a post by Sinders [23]) have highlighted such concerns. A flaw of these detectors is how differently they flag language used by or about particular groups of people.

Our goal in this paper is to address discrimination against particular groups of people in the context of cyberbullying detection. Many machine learning algorithms have been introduced to detect cyberbullying in social media. Raisi & Huang in [21] introduced a framework called *co-trained ensembles*, which uses weak supervision to significantly alleviate the need for tedious data annotation. Their weak supervision is in the form of expert-provided key phrases that are highly indicative or counter-indicative of bullying. In addition, their framework is based on consistency of two detectors that co-train one another. These detectors use two different perspectives of the data: (1) language and (2) social structure. By using different forms of evidence, the detectors train to reach the same conclusion about whether social interactions are bullying. Furthermore, they incorporate distributed word and graph-node representations by training nonlinear deep models.

With the advantages of weakly supervised training, there is also a concern that the self-training mechanism used to amplify the weak supervision may also amplify patterns of societal bias. Therefore, in this paper, we extend the co-trained ensemble model to mitigate unfair behavior in the trained model. We add unfairness penalties to the training framework introduced by Raisi & Huang [21] when we observe discrimination in predictions. We explore two unfairness penalty terms, each aiming toward a different notion of fairness. One targets *removal fairness* and the other targets *substitutional fairness*. For removal fairness, we penalize the model if the score of a message containing sensitive keywords is higher than if those keywords were removed. For substitutional fairness, for each protected group, we provide a list of sensitive keywords and appropriate substitutions. For example, for the keyword "black" describes an ethnicity, and substitutions are "asian," "american," "middle-eastern," etc. A fair model would score a message containing any sensitive keyword the same if we replace that sensitive keyword with another; otherwise, we penalize the objective function.

We measure the learned model's fairness on synthetic data and a dataset of Twitter messages. Our synthetic data is a corpus of sentences generated using the combination of some sensitive keywords describing different attributes: sexual orientation, race, gender, and religion. Mirroring the benchmark established by Raisi & Huang [20], we generated statements of identity (e.g., "black woman", "muslim middle-eastern man") that are not harassment. To assess each model's fairness, we compute the *false-positive rate* on these identity statements. Since these statements are not bullying, an ideal fair language-based detector should yield a lower false-positive rate on these examples. For evaluation on Twitter data, we measure model fairness using the equality of odds gaps [10]. Specifically, we use a criterion we call *category dispersion*, which is the standard deviation of area under the curve (AUC) of receiver operating characteristic (ROC) curves across multiple keywords in a category of substitutable keywords. A low category dispersion is more desirable since it indicates that a model treats the subpopulations equitably. A high category dispersion indicates that the model behavior is more

favorable toward some subgroups; hence, it discriminates against some other subpopulations. We also test each model's fairness qualitatively by examining conversations with sensitive keywords where their score using the reduced-bias model is much lower than the default model. In another qualitative analysis, we examine the change in the bullying score of sensitive keywords when fairness imposed to the harassment detector. Together, our evaluations demonstrate the ability of our approach to reduce the biases of weakly supervised bullying detectors.

2 Related Work

This study mainly builds on two bodies of research: (1) machine learning for detection of online harassment and cyberbullying, and (2) fairness in machine learning. We cover only the most directly relevant literature because of limited space.

Online Harassment and Cyberbullying Detection A variety of methods have been proposed for cyberbullying detection. These methods mostly approach the problem by treating it as a classification task, where messages are independently classified as bullying or not. Many of the research contributions in this space involve the specialized design of message features for supervised learning. Many contributions consider features based on known topics used in bullying [5,7,22], sentiment [27], topic models [17], vulgar language expansion [19], audio and video features [24], and social structure features [3,14,30,4]. Hosseinmardi et al. studied negative user behavior in the Ask.fm and Instagram [11,12]. Tomkins et al. propose a probabilistic model [25] for cyberbullying detection. Raisi and Huang [21,20] introduced a weakly supervised machine learning method for cyberbullying detection. Our work builds on their approach.

Fairness in Machine Learning In recent years, there has been rapid progress in designing *fair* machine learning algorithms. Machine learning algorithms can exhibit discriminatory decision making in areas such as recommendation, prediction, and classification [1,15]. Various fairness measures have been introduced such as *equal opportunity* [10] and *disparate mistreatment* [28]. Zhang et al. [29] examine three fairness measures: demographic parity, equality of odds, and equality of opportunity in the context of adversarial debiasing. Garg et al. [8] introduce counterfactual fairness in text classification by substituting individual tokens related to sensitive attributes. This approach is similar to ours in this paper, but our method is aimed toward weakly supervised learning for cyberbullying detection.

3 Review of Co-trained Ensembles for Weak Supervision

In this section, we review the co-trained ensemble framework introduced by Raisi & Huang [21] and how it is applied to train cyberbullying detectors. The approach learns from weak supervision by seeking consensus between two model families: 1) message classifiers and 2) user classifiers. Message classifiers take a message as input and output a classification score for whether the message is an example of harassment. User classifiers take an ordered pair of users as input and output a score indicating whether one

user is harassing the other user. For message classifiers, the framework accommodates a generalized form of weakly supervised loss function ℓ (which could be extended to also allow full or partial supervision). Let Θ be the model parameters for the combined ensemble of both classifiers. The training objective is

$$\min_{\Theta} \underbrace{\frac{1}{2|M|}\sum_{m\in M}\left(f(m;\Theta)-g\left(s(m),r(m);\Theta\right)\right)^2}_{\text{consistency loss}}+\underbrace{\frac{1}{|M|}\sum_{m\in M}\ell\left(f(m;\Theta)\right)}_{\text{weak supervision loss}} \quad (1)$$

where M is a set of all messages and $s(m)$ and $r(m)$ are the sender and receiver of message m. The first loss function is a consistency loss, and the second loss function is the weak supervision loss. The consistency loss penalizes the disagreement between the scores output by the message classifier for each message and the user classifier for the sender and receiver of the message. The weak supervision relies on annotated lists of key-phrases that are indicative or counter-indicative of harassment. Let there be a set of indicator phrases and a set of counter-indicator phrases for harassment. The weak supervision loss ℓ is based on the fraction of indicators and counter-indicators in each message, so for a message containing $n(m)$ total key-phrases, let $n^+(m)$ denote the number of indicator phrases in message m and $n^-(m)$ denote the number of counter-indicator phrases in the message. The weak supervision loss is

$$\ell(y_m) = -\log\left(\min\left\{1, 1+(1-\tfrac{n^-(m)}{n(m)})-y_m\right\}\right)-\log\left(\min\left\{1, 1+y_m-\tfrac{n^+(m)}{n(m)}\right\}\right). \quad (2)$$

Within this framework, they used classifiers built on vector representations of message and users. Vector embeddings of text are incorporated into the framework in two ways: 1) using existing word-embedding models as inputs to message classifier, 2) creating new embedding models specifically geared for analysis of cyberbullying. The user classifiers represent each user as a vector, classifying the vector pair as either a bullying or a non-bullying relationship.

Raisi & Huang [21] examined four message classifiers: (1) BoW: a randomly hashed bag of n-grams model with 1,000 hash functions [26], (2) doc2vec: a linear classifier based on the pre-trained doc2vec vector of messages trained on the dataset [16], (3) emb: a custom-trained embedding model with each word represented with 100 dimensions, (4) RNN: a recurrent neural network with two hidden layers of dimensionality 100. The emb and RNN models are trained end-to-end to optimize overall loss function, while pre-trained models (BoW, doc2vec) are used to only adjust the linear classifier weights given the fixed vector representations for each message. For the user classifiers, they use a linear classifier on concatenated vector representations of the sender and receiver nodes. To compute the user vector representations, they pre-train a node2vec [9] representation of the communication graph. The pre-trained user vectors are then the input to a linear classifier that is trained during the learning process. There are eight combinations of message and user learners (including the option to use no user learner, in which case we use the weak supervision loss to train message classifiers).

According to the quantitative experimental evaluation in [21], the top five combinations of text and user classifiers that produce the highest precision@100 **on Twitter** are: "emb_none," "emb_node2vec," "bow_none,", "rnn_none," and "doc2vec_node2vec." In our experiments, we do not consider "rnn_none" because its performance was similar

to "bow_none" but was much more computationally expensive. Therefore, we use four configurations of their ensemble framework for our experiments.

4 Reduced-Bias Co-Trained Ensembles

In this section, we introduce our approach to reduce bias in co-trained ensemble models. Our goal is to disregard discriminatory biases in the training data and create fair models for cyberbullying detection. We add a term to the loss function to penalize the model when we observe discrimination in the predictions. We investigate the effectiveness of two such *unfairness penalty terms*.

Removal Fairness The motivation for *removal fairness* is that, in a fair model, the score of a message containing sensitive keywords should not be higher than the same sentence without these keywords. Therefore, we penalize the model with an unfairness loss:

$$\ell(y_m) = \alpha \times -\log\left(\min\left\{1, 1 - (y_m - y_{m^-})\right\}\right). \tag{3}$$

where y_m is the score of message containing sensitive keywords and y_{m^-} is the score of the same message when sensitive keywords are dropped. The parameter α represents to what extent we enforce fairness to the model. In our experiments, we examined three values $\alpha \in \{1, 10, 100\}$. The best value resulting better generalization and lower validation error was $\alpha = 10$.

Substitutional Fairness In *substitutional fairness*, for each category of sensitive keyword, we provide a list of sensitive keywords and appropriate substitutions. For example, the keyword "black" describes a type of ethnicity, so substitutions are "asian," "native-american," "middle-eastern," etc. Or the keyword "gay" describes a sexual orientation, so it can be substituted with "straight," "bisexual," "bi," etc. In a fair model, the score of a message containing a sensitive keyword should not change if we replace that sensitive keyword with its substitutions. We penalize the objective function with

$$\ell(y_m) = \frac{\alpha}{|S_c| - 1} \times \sum_{i \in S_c, i \neq k} \left(y_{m(k)} - y_{m(i)}\right)^2. \tag{4}$$

where S_c is the set of all sensitive keywords in category c, $|S_c|$ is the cardinality of set S_c, $y_{m(k)}$ is the score of original sentence containing sensitive keyword k in category c, and $y_{m(i)}$ is the score of the substitution sentence with sensitive keyword i in category c. As with removal fairness, α represents the strength of the unfairness penalty. We tested $\alpha \in \{1, 10, 100\}$, and the value 10 again led to the best validation error.

5 Experiments

Mirroring the setup initially used to evaluate the co-trained ensemble framework, we construct a weak supervision signal by collecting a dictionary of offensive key-phrases (unigrams and bigrams) [18]. For our weak supervision, we manually curated a collection of 516 high-precision bullying key phrases that do not specifically target particular groups. We augment this with a list of positive opinion words collected by Hu & Liu [13]. The offensive phrases are our weak indicators and the positive words are our

counter-indicators. Out of eight combinations of message and user learners introduced in [21], we selected the top four models with the highest precision@100 on Twitter for our evaluation: "emb_none," "emb_node2vec," "bow_none," and "doc2vec_node2vec."

We consider four categories of sensitive keywords: race, gender, religion, and sexual orientation. For each category, we provide a list of keywords. Some example keywords in the race category are "white," "black," "caucasian," "asian," "indian," and "latina"; and some example keywords in religion category are "christian," "jewish," "muslim," "mormon," and "hindu." In each message, there might be many sensitive keywords. Hence, for computational purposes, we limit the number of substitutions to a randomly selected 20 substituted sentences.

Evaluation on Synthetic Data We analyze the sensitivity of models toward some often targeted groups on a synthetic benchmark established by [20]. This benchmark is a corpus of sentences using the combination of sensitive keywords describing different attributes. These statements of identity (e.g., "I am a Black woman," "I am a Muslim middle-eastern man," etc.) are not bullying since they are simply stating one's identity. In total, there are 2,777 non-bullying statements. To assess each model's fairness, we compute the *false-positive rate* on these synthetic benchmark statements. An ideal fair language-based detector should yield a lower false-positive rate on these non-toxic statements. Table 1 shows the *false-positive rates* (at threshold 0.5) of four aforementioned co-trained ensembles with and without penalizing unfairness. Since the generated statements are not bullying, the false-positive rate of an ideal fair model should be 0.0. According to the results in Table 1, the false-positive rate of these four methods reduces when either removal or substitutional fairness constraints applied. When removal fairness is imposed to the "emb_none" and "bow_none", their false-positive rate reduced to zero. The false-positive rate of "doc2vec_node2vec" with and without enforcing fairness is zero. Since the synthetic data does not have social networks to train node2vec, we just used the message learner to give bullying score to these statements.

Table 1: False positive rate of models on non-bullying synthetic benchmark statements (using threshold 0.5). Both *removal* and *substitutional* fair models reduce the false positive rate compare to without bias reduction (vanilla).

Method	emb_none	emb_node2vec	bow_none	doc2vec_node2vec
Vanilla	0.8416	0.7055	0.2663	0.0000
Substitutional	0.7685	0.1439	0.0305	0.0000
Removal	0.0000	0.0418	0.0000	0.0000

Evaluation on Twitter We use the data collected by Raisi & Huang [21]. They collected data from Twitter's public API, extracting tweets containing offensive-language words posted between November 1, 2015, and December 14, 2015. They then extracted conversations and reply chains that included these tweets. They then used snowball sampling to gather tweets in a wide range of topics. After some preprocessing, the Twitter data contains 180,355 users and 296,308 tweets.

We evaluate the effectiveness of our approach using post-hoc crowdsourced annotation to analyze the score of fair-imposed model for conversations with sensitive keywords. We extract all of the conversations in Twitter data containing at least one sensitive keyword. We then asked crowdsource workers from Amazon Mechanical Turk to evaluate the interactions. We showed the annotators the anonymized conversations and asked them, "Do you think either user 1 or user 2 is harassing the other?" The annotators indicated either "yes," "no," or "uncertain." We collected three annotations per conversation. For each of the three annotators, we score a positive response as +1, a negative response as -1, and an uncertain response as 0. We sum these scores for each interaction, and we consider the interaction to be harassment if the score is 2 or greater.

To measure a model's fairness, we use an "equality of odds" criterion, which states all subpopulations in a group experience the same true- or false-positive rate [10]. We generalize a notation "accuracy equity" [6]. We compute the standard deviation of the area under the curve (AUC) of the receiver order characteristic (ROC) for messages containing keywords in a category of sensitive keywords. We refer to this measure as *category dispersion*. For example, in the "religion" category, we compute the category dispersion across different keywords in this category such as "muslim," "christian," "jewish," "protestant," etc. An ideal, fair language-based detector should treat these keywords equitably, which would induce a lower category dispersion.

Table 2 shows the category dispersions of methods on each targeted group. We bold values when the category dispersion of the reduced-bias method is lower than the vanilla learner (without fairness), which indicates that the reduced-biased model treats the keywords in the protected category more equitably. Enforcing substitutional fairness on emb_none and doc2vec_node2vec leads to fairer models across all three categories. Enforcing substitutional fairness on bow_none leads to fairer models in two out of three categories; and substitutional emb_node2vec is fairer for only the religion category. Enforcing removal fairness on emb_none, doc2vec_node2vec, and bow_none leads to fairer behavior in two out of three categories, while doing so on emb_node2vec only treats keywords in the religion category more equitably. In summary, emb_none and doc2vec_node2vec, when trained with substitutional fairness terms produce lower standard deviation of AUC across keywords for all three tested categories.

An important question is whether there is accuracy degradation as our approach encourages more equitable errors within categories. In Figure 1, we plot the ROC curve of four co-trained ensemble methods stratified by fairness type. Surprisingly, the AUC of substitutional bow_none, emb_none, and doc2vec_node2vec actually improve over the default approach. The performance of removal emb_none improves over vanilla, but other methods' performance reduce when adding removal fairness. In Figure 2 we compare the ROC curve of emb_none method for some sensitive keywords for the categories of sexual orientation and religion. The ROC curves of sensitive keywords in each group are closer to each other in both removal and substitutional fair methods, while the AUC of most keywords in the substitutional and removal versions are higher than the vanilla approach. This trend indicates that bias-reduction successfully equalizes behavior across language describing different subpopulations of people.

Qualitative Analysis We qualitatively test the model's fairness by analyzing the highest scoring conversations identified by the models. An ideal fair model should give a

Table 2: The category dispersion of four co-trained ensemble methods for three targeted groups. A bold value means the category dispersion of the reduced-bias model is lower than the default method (vanilla). Substitutional emb_none and doc2vec_node2vec have better category dispersion than the vanilla method in all three groups.

Category	Fairness Type	emb_none	emb_node2vec	bow_none	doc2vec_node2vec
Race	Vanilla	0.0937	0.0525	0.0231	0.0546
	Substitutional	**0.0531**	0.0528	0.0514	**0.0460**
	Removal	**0.0640**	0.0594	0.0665	0.0587
Religion	Vanilla	0.0858	0.0862	0.0716	0.0748
	Substitutional	**0.0657**	**0.0376**	**0.0660**	**0.0494**
	Removal	0.0902	**0.0424**	**0.0190**	**0.0661**
Sexual Orientation	Vanilla	0.1096	0.0791	0.0915	0.0702
	Substitutional	**0.0629**	0.0821	**0.0666**	**0.0623**
	Removal	**0.0894**	0.0914	**0.0467**	0.0784

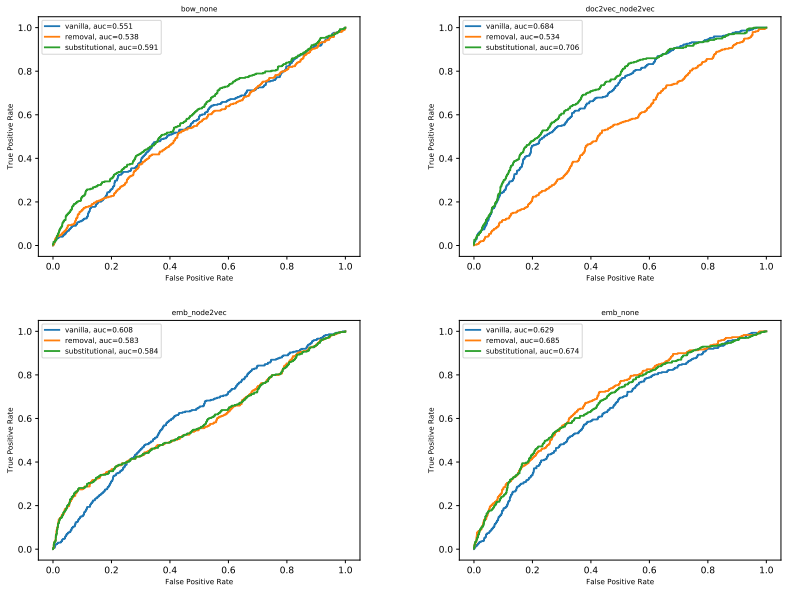
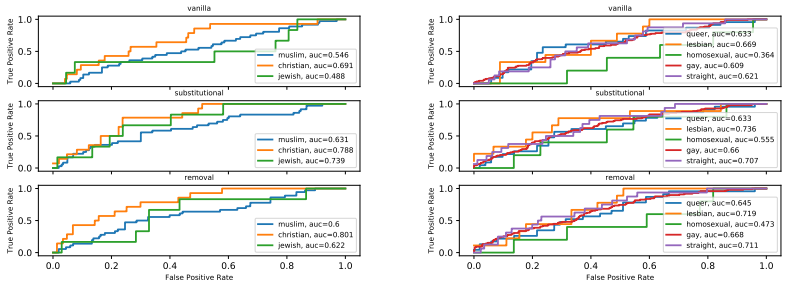


Fig. 1: ROC curve of the four co-trained ensemble methods stratified by fairness type. The performance of substitutional bow_none, doc2vec_node2vec, and emb_none improves over the vanilla method. Removal emb_none also improves over vanilla, but emb_node2vec has worse performance.

lower score to non-bullying interactions containing sensitive keywords. Figure 3 displays three non-bullying conversations highly ranked by the vanilla model, but given a low score by the reduced-bias model.

We also compare the bullying score of messages containing only one sensitive keyword with and without bias reduction. We consider the scores of embedding message classifiers since they learned a vector representation for all words in corpus. We plot the



(a) emb_none for religion (b) emb_none for sexual orientation

Fig. 2: ROC curves for the emb_none model on data containing sensitive keywords in the sexual orientation and religion categories. The gap between AUC of sensitive keywords in these two groups reduces, without degradation in performance.

User1: all figure that you are, in fact, ***gay*** and theyre freaking out about it? F*CK THOSE PEOPLE.
User2: I think my Uncle has figured it out, him & my cuz are always trying to out me, I feel nervous/sick every time theyre around.

User1: Alienate minorities? Why would anyone, of any color, stand w/ppl that dont believe al lives matter?
User2: All Lives matter Is a gasp of a little racist mind, when Cops shot white boys at racist of ***Blacks*** then OK

User1: you cant really be racist against a religion...
User2: i know my friend but you can be racist against followers of this religion.iam an exMuslim but refuse to be a racist against them
User1: Racist is hating someone based in their skin color, not their faith. Hating someone just for their faith is just bigotry.
User2: remamber the ***Jewish*** and the holocaust or you think it wasnt racist!!!!!!!!!!!! #ExMuslimBecause
User1: ***Jewish*** can be a race or a religion. There is no Muslim race. Theres an Arab race, but you cant assume Arabs are Muslim.

Fig. 3: Three examples of conversations given a high score by the vanilla model and given a low score by a reduced-bias model. These conversations discuss sensitive topics but may not represent one user bullying another.

scores in Figure 4. Our findings are as follows: The score of most sensitive keywords in the ethnicity category, such as “black,” “indian,” “african,” “american,” “asian,” “hispanic,” and “white” reduces when fairness is imposed on the model. The reason the keyword “white” is scored higher using the vanilla model could be the occurrence of this word in bullying interactions in our dataset. In the gender category, the score of “boy” and “man” increases when imposing substitutional fairness, but the score of “girl” and “woman” either reduces or does not change. In the religion category, the bullying score of “muslim,” “protestant,” “mormon,” and “jewish” reduce when fairness is imposed. In the sexual orientation category, the score of “transexual,” “queer,” “lesbian,” “bi,” “heterosexual,” “homosexual,” and “straight” reduce using biased-reduced models.

One interesting observation is how well the score of message with sensitive keywords becomes more uniform after enforcing the fairness to the model. More particularly, in the emb_node2vec model, the score of two keywords “boy” and “girl” get closer to each other after imposing substitutional fairness. This is also true for “woman” and

"man." By scrutinizing the behavior of the emb_none model, we observe a remarkable change in the score of sensitive keywords when fairness is imposed on the model. Its vanilla version gives the highest bullying score to most sensitive keywords, but the scores reduce noticeably when fairness is imposed. On the other hand, the score of some keywords using the vanilla method is significantly low, but when fairness terms are applied, their scores increase. An open question is whether this significant variation is desirable especially with considering the performance improvement of emb_none in our quantitative analysis.

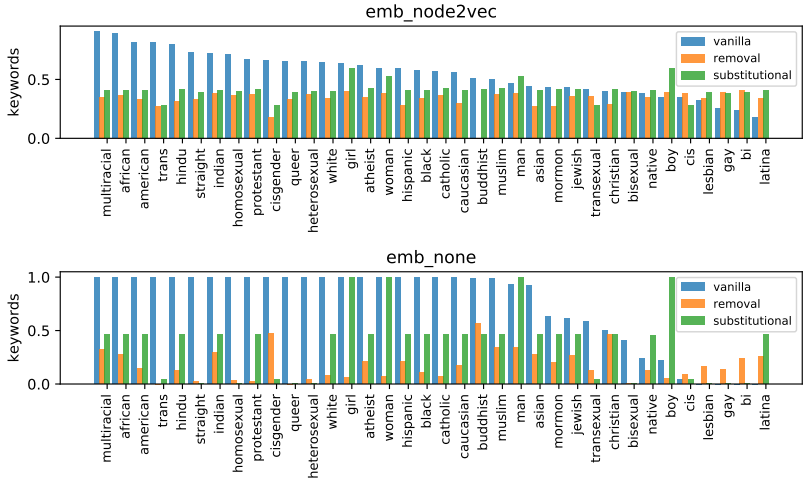


Fig. 4: Bullying scores of messages containing only one sensitive keyword as scored by emb_none and emb_node2vec with and without fairness terms. The bullying score of most sensitive keywords reduce when fairness is imposed on the model, and the scores become more uniform than without any fairness terms (vanilla).

Discussion One question that might arise is which fairness objective best provides the desired fairness? Considering our quantitative analysis in Table 2, which follows the equality of odds criterion, substitutional constraints improve the AUC gap of more groups for each method. However, the difference is not significant. Both penalty terms reduce the false-positive rate on the synthetic benchmark. Another question is which combination of model architecture and fairness objective has better behavior? Table 2 suggests emb_none and doc2vec_node2vec with substitutional fairness a produce lower AUC gap between keywords for all three groups. One might ask about the trade-off between accuracy and fairness. As shown in Figure 2, adding a fairness term reduces the accuracy of emb_node2vec, but emb_node2vec is also unable to produce fair predictions either. This pattern suggests that emb_node2vec is not compatible with the introduced fairness constraints. The emb_none and doc2vec_node2vec models have better fairness

behavior with the substitutional objective, and their accuracy also improves. The removal objective, however, reduces the accuracy when added to most models.

6 Conclusion

Fairness is one of the most important challenges for automated cyberbullying detection. As researchers develop machine learning approaches to detect cyberbullying, it is critical to ensure these methods are not reflecting or amplifying discriminatory biases. In this paper, we introduce a method for training a less biased machine learning model for cyberbullying analysis. We add unfairness penalties to the learning objective function to penalize the model when we observe discrimination in the model's predictions. We introduce two fairness penalty terms based on removal and substitutional fairness. We use these fairness terms to augment co-trained ensembles, a weakly supervised learning framework [21]. We evaluate our approach on a synthetic benchmark and real data from Twitter. Our experiments on the synthetic benchmark show lower *false-positive rates* when fairness is imposed on the model. To quantitatively evaluate model's fairness on Twitter, we use an equality of odds measure that computes the standard deviation of AUC for messages containing sensitive keywords in a category. A fair model should treat all keywords in each category equitably, i.e., have a lower standard deviation. We observe that two ensemble learners, when augmented with substitutional fairness, reduce the gap between keywords in three groups, while their detection performance actually improves. We did not always observe such behavior when models were augmented with removal fairness. In addition, we qualitatively evaluate the framework, extracting conversations highly scored by the vanilla model but not flagged by the bias-reduced models. These conversations tended to be false-positive, non-bullying conversations that used sensitive language. We therefore demonstrate the capability to reduce unfairness in cyberbullying detectors trained with weak supervision.

References

1. Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. CoRR **abs/1607.06520** (2016)
2. Boyd, D.: It's Complicated. Yale University Press (2014)
3. Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E.D., Stringhini, G., Vakali, A.: Mean birds: Detecting aggression and bullying on twitter. CoRR **abs/1702.06877** (2017)
4. Chelmis, C., Zois, D.S., Yao, M.: Mining patterns of cyberbullying on twitter. 2017 IEEE International Conference on Data Mining Workshops (ICDMW) pp. 126–133 (2017)
5. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. Intl. Conf. on Social Computing pp. 71–80 (2012)
6. Dieterich, W., Mendoza, C., Brennan, T.: Compas risk scales : Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county (2016)
7. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. ICWSM Workshop on Social Mobile Web (2011)
8. Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E.H., Beutel, A.: Counterfactual fairness in text classification through robustness. CoRR **abs/1809.10610** (2018)

9. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. CoRR **abs/1607.00653** (2016)
10. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. CoRR **abs/1610.02413** (2016)
11. Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., Mishra, S.: Towards understanding cyberbullying behavior in a semi-anonymous social network. Intl. Conf. on Adv. in Social Networks Analysis and Mining pp. 244–252 (2014)
12. Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q., Mishra, S.: Detection of cyberbullying incidents on the Instagram social network. Association for the Advancement of Artificial Intelligence (2015)
13. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proc. of the ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining. pp. 168–177 (2004)
14. Huang, Q., Singh, V.K.: Cyber bullying detection using social and textual analysis. Proceedings of the International Workshop on Socially-Aware Multimedia pp. 3–6 (2014)
15. Kim, M.P., Ghorbani, A., Zou, J.Y.: Multiaccuracy: Black-box post-processing for fairness in classification. CoRR **abs/1805.12317** (2018)
16. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proc. of the International Conference on Machine Learning. pp. 1188–1196 (2014)
17. Nahar, V., Li, X., Pang, C.: An effective approach for cyberbullying detection. Communications in Information Science and Management Engineering **3**(5), 238–247 (May 2013)
18. noswearing.com: List of swear words & curse words. http://www.noswearing.com (2016)
19. Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., Araki, K.: Machine learning and affect analysis against cyber-bullying. In: Linguistic and Cognitive Approaches to Dialog Agents Symposium. pp. 7–16 (2010)
20. Raisi, E., Huang, B.: Co-trained ensemble models for weakly supervised cyberbullying detection. In: NeurIPS Workshop on Learning with Limited Labeled Data (2017)
21. Raisi, E., Huang, B.: Weakly supervised cyberbullying detection using co-trained ensembles of embedding models. In: Proc. of the IEEE/ACM International Conference on Social Networks Analysis and Mining. pp. 479–486 (2018)
22. Rezvan, M., Shekarpour, S., Thirunarayan, K., Shalin, V.L., Sheth, A.P.: Analyzing and learning the language for different types of harassment. CoRR **abs/1811.00644** (2018)
23. Sinders, C.: Toxicity and tone are not the same thing: analyzing the new Google API on toxicity, PerspectiveAPI (2017), `https://medium.com/@carolinesinders/toxicity-and-tone-are-not-the-same-thing-analyzing-the-new-google-api-on-toxicity-perspectiveapi-14abe4e728b3`
24. Soni, D., Singh, V.K.: See no evil, hear no evil: Audio-visual-textual cyberbullying detection. Proc. ACM Hum.-Comput. Interact. **2**, 164:1–164:26 (2018)
25. Tomkins, S., Getoor, L., Chen, Y., Zhang, Y.: A socio-linguistic model for cyberbullying detection. In: Intl. Conf. on Advances in Social Networks Analysis and Mining (2018)
26. Weinberger, K., Dasgupta, A., Langford, J., Smola, A., Attenberg, J.: Feature hashing for large scale multitask learning. In: Proc. of the Intl. Conf. on Machine Learning. pp. 1113–1120 (2009)
27. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on Web 2.0. Content Analysis in the WEB 2.0 (2009)
28. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proc. of the Intl. Conf. on World Wide Web. pp. 1171–1180 (2017)
29. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. CoRR **abs/1801.07593** (2018)
30. Zois, D.S., Kapodistria, A., Yao, M., Chelmis, C.: Optimal online cyberbullying detection. 2018 IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc. pp. 2017–2021 (2018)