A Fairness-Aware Fusion Framework for Multimodal Cyberbullying Detection

Jamal Alasadi
Rutgers University, USA
University of Thiqar, Iraq
jamal.alasadi@rutgers.edu

Ramanathan Arunachalam
Rutgers University, USA
ramanathan.arun@rutgers.edu

Pradeep K. Atrey
University at Albany,
State University of New York, USA
patrey@albany.edu

Vivek K. Singh
Rutgers University, USA
v.singh@rutgers.edu

***Abstract*—Recent reports of bias in multimedia algorithms (e.g., lesser accuracy of face detection for women and persons of color) have underscored the urgent need to devise approaches which work equally well for different demographic groups. Hence, we posit that ensuring fairness in multimodal cyberbullying detectors (e.g., equal performance irrespective of the gender of the victim) is an important research challenge. We propose a fairness-aware fusion framework that ensures that both fairness and accuracy remain important considerations when combining data coming from multiple modalities. In this Bayesian fusion framework, the inputs coming from different modalities are combined in a way that is cognizant of the different confidence levels associated with each feature and the interdependencies between features. Specifically, this framework assigns weights to different modalities not just based on accuracy but also their fairness. Results of applying the framework on a multimodal (visual + text) cyberbullying detection problem demonstrate the value of the proposed framework in ensuring both accuracy and fairness.**

***Keywords*-Cyberbullying Detection, Fairness, Bias in Machine Learning, Multimedia Fusion, Bayesian Fusion**

I. Introduction

Cyberbullying is an increasingly complicated problem faced by many online social networks users. As stated in [1], cyberbullying occurs "when the Internet, cellphones or other devices are used to send or post text or images intended to hurt or embarrass another person". The National Crime Prevention Council reports more than 40% of teenagers in the US have reported being cyberbullied [2]. When dealing with large-scale social networks, it is impractical to use a completely manual approach for cyberbullying detection. Hence, multiple researchers have proposed machine learning-based methods for automatic cyberbullying detection. While past research on cyberbullying detection has been dominated by text mining and analysis [3], [4], multiple authors have started using multimodal content analysis for cyberbullying detection. Recent attempts have recognized the value of looking at the heterogeneous modalities of information including textual features, social features, audio and visual features for improved cyberbullying detection [5], [6], [7], [8].

Like many fields, the adoption of machine learning approaches has led to significant advancements (e.g. scalability, accuracy) in cyberbullying detection. At the same time, the issue of fairness in machine learning algorithms has gained prominence. Multiple authors have criticized multimedia algorithms that perform differently for different groups [9] and some recent efforts have studied the issue of fairness in cyberbullying detection [10], [11].

While past research in multimedia has used the multiple perspectives coming from different modalities to improve on accuracy of the developed algorithm, we posit that *combining input from multiple modalities is also a natural way to reduce bias*. In this paper we consider the problem of fairness in multimodal cyberbullying detection and make the following contributions:

1) Undertake the first audit on fairness of multimodal cyberbullying detection algorithms.
2) Propose a fairness-aware Bayesian fusion framework for multimodal cyberbullying detection.

The underlying intuition behind the framework is as follows. Each modality captures a part of the underlying phenomena and has some inherent predictive power but also some inherent bias. Further, there are varying levels of interdependencies between modalities. Hence, the framework needs to combine the inherent bias levels, accuracy levels, and interdependencies in a manner that optimizes for a combination of accuracy and fairness.

In this work, we focus on a human-labeled cyberbullying data-set of Instagram "media sessions" (image and textual captions and comments from other users), which has been used in a past works [12], [13]. The dataset includes information on the gender of the person in the image as obtained via a computer vision API. We operationalize bias as the difference in the performance of the model for privileged vs. unprivileged (here, male vs. non-male) groups. The results obtained suggest that the proposed framework yields improvements in fairness while maintaining a high degree of accuracy.

The proposed work has limitations including the use of binary gender labels based on visual presentation [14]. However, since reaching out to the original victims of cyberbullying to obtain self-reports on gender is neither practical nor ethical, we have decided to use the abovementioned dataset to test out the proposed multimodal fairness aware framework. We note that the proposed framework is generic and works for other interpretations of accuracy, bias, and interdependencies too.

978-1-7281-9325-0/20/$31.00 ©2020 IEEE
DOI 10.1109/BigMM50055.2020.00032

The rest of this paper is organized as follows. In Section II, we cover related work in more details. Section III, discusses some concepts of fairness in machine learning and the metrics. Section IV, explains the proposed framework and describes the scenarios that we consider in our study. Implementation details and dataset used in this work are described in Section V. Our evaluation results are presented in Section VI, Subsequently, Section VII, concludes the paper and provides future directions.

II. Related Work

A. Cyberbullying Detection

Text-based analysis remains an important area of cyberbullying detection research. In an early attempt, Reynolds et al. in [4] used the amount, density, and value of swear words as sensitive features to detect the cyberbullying messages. Huang et al. in [6] developed an approach to utilize social and textual features in a combined cyberbullying detections. Dinakar et al. in [15] proposed a model that showed that label-particular classifiers are more efficient than multi-class classifiers in detecting cyberbullying messages. Singh et al. have defined a fusion framework for cyberbullying detection using social and textual features [5]. Recently, multiple authors have stated using multimodal signals i.e. audio, visual, textual, and social features for advanced cyberbullying detection [2], [16], [17]. Some recent attempts have utilized embeddings and deep learning methods to create advanced text-based cyberbullying detectors [18], [19]. While deep learning approaches often yield high accuracy, they also require large-scale labeled data. However, the nuances of cyberbullying often require human involvement in the labeling process and large-scale datasets for training such models are available only in limited settings [20].

B. Fairness in Machine Learning

In recent years, there have been multiple attempts that target improving the fairness of machine learning (ML) algorithms. Fairness in ML algorithms is typically operationalized as equal performance of the algorithm for different demographic groups [9]. If on the other, the performance of the algorithm varies depending on a "sensitive" attribute (e.g., race, gender) then the algorithm is considered biased i.e. not fair [21].

Zafar et al. [22] described a framework for achieving the middle ground between fairness and accuracy. Recent studies have shown that employing Generative Adversarial Networks, where one convolution neural network optimizes for accuracy and other for bias can be useful for ensuring fairness and accuracy in machine learning applications. For instance, Alasadi et al. [23] implemented a convolution neural network adversary that finds the best accuracy for face matching while reducing discrepancy in TPR (true positive rate) and FPR (false positive rate) across different demographic descriptors. Bechavod et al. [24] proposed an approach for reducing unfairness classification using logistic regression on binary classification tasks over persons from two groups of populations in applications like criminal risk assessment and college admissions. Amini et al. [25] reduced the bias by proposing a method to modify the sampling probabilities of independent data points while training an algorithm. Consequently, reducing hidden biases entirely within the training dataset, the model showed increased accuracy and fairness. While multiple similar efforts for improving fairness have been emerging in recent literature, the issue of fairness in multimodal fusion remains an open challenge.

C. Multimodal Fusion Approaches

Information fusion is an important sub-area of multimedia research [1], [26], [27]. Atrey et al. [28] proposed a system to detect the important events in multimedia surveillance system that uses a hierarchical probabilistic method to identify events and combined them in the assimilation process that helps in achieving the overall accuracy of event detection. Wang et al., [26] describe an approach fusing information coming from multiple sensors across space and time based on a unified "cmage" representation. Kong et al. [27] describe an approach where the correlation among multiple sensors is used to fuse the monitoring information and improve the resolution and accuracy of the system. In essence, while improving accuracy and coverage has often been a goal of multimedia fusion approaches, the goal of improving fairness based on multimodal fusion is yet to be explored.

III. Preliminaries: Fairness in Machine Learning

Let us consider a classical supervised classification problem trained with m examples $(x_i, s_i, y_i)_{i=1}^{m}$ where x_i is a feature vector with P predictors of the i^{th} example, s_i is its sensitive attribute and y_i is a label. Here the prediction of the algorithm for the i^{th} example is represented as $\hat{y}_i$. In order to realize fairness, it is important to have a clear understanding of its formal definition. In the following, we summarize the most popular definitions used in recent Fairness in Machine Learning research.

First, there is data sanitization which concerns the information that is used for training the model (e.g., race should not be used as a feature for recidivism prediction). Second, there is individual fairness, which relates at the individual level and proposes that fairness means similar individuals must be treated similarly. Finally, there is group fairness, which is a type of fairness that divides the world into groups defined by one or multiple high-level sensitive attributes. It needs a particular relevant statistic (e.g., accuracy, true positive rate) about the classifier to be the same across those combinations. We concentrate on this family of fairness measures and describe the popular definitions of these kinds used in recent research [29]. In one such definition, a classifier is

considered to make a fair decision if the prediction $\hat{Y}$ from features X is independent of the protected attributes S (e.g., gender) [21] i.e.

$$P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1) \quad (1)$$

Such absolute notion of fairness is rarely achieved in practical systems. We discuss some of the other common metrics for fairness below.

A. Demographic Parity

The underlying idea is that each demographic group should have the same opportunity for a positive result. There are several ways to evaluate this objective. The P-rule calculation ensures the ratio of positive rate for the unprivileged set is no less than a specific threshold (e.g., τ =80%) and is given by:

$$min(\frac{P(\hat{Y} = 1|S = 1)}{P(\hat{Y} = 1|S = 0)}, \frac{P(\hat{Y} = 1|S = 0)}{P(\hat{Y} = 1|S = 1)}) \geq \tau \quad (2)$$

The classifier is considered as totally fair if this ratio is 100%; a 0% score indicates a completely unfair model. Another measurement which can be used for demographic parity is the disparate impact (DI) valuation [30]. It takes into consideration the absolute difference of result distributions for sub-populations with various sensitive features values. The value of DI can be calculated as:

$$DI = |P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S = 0)| \quad (3)$$

A smaller value of DI indicates a fairer model.

B. Equalized Odds

A model is considered fair when across both demographic groups ($S = 0$ and $S = 1$), the predictor $\hat{y}$ has equal TPR and equal FPR [31]. This enforces that the accuracy is equally high for particular sub-populations within the overall population because of the rate of positive and negative classification is the same across such groups.

$$P(\hat{Y} = 1|Y = y, S = 0) = P(\hat{Y} = 1|Y = y, S = 1) \quad (4)$$

where $y \in (0, 1)$.

This objective can be measured by a metric that determines the disparate mistreatment [22]. It calculates the total differences between TPRs and FPRs for both demographics, given by D_{TPR} and D_{FPR}, respectively, which are computed as follows:

$$D_{TPR} : |P(\hat{Y} = 1|Y = 1, S = 1) - P(\hat{Y} = 1|Y = 1, S = 0)| \quad (5)$$

$$D_{FPR} : |P(\hat{Y} = 1|Y = 0, S = 1) - P(\hat{Y} = 1|Y = 0, S = 0)| \quad (6)$$

IV. Proposed Approach

This work proposes a method to improve the performance of cyberbullying detection methods beyond that is obtainable by using a single modality of data, or its 'naive' assimilation. Naive assimilation here refers to the most common approach of simply combining the features as-is, without identifying the different confidence values for different features or the inter-dependencies between them.

A. Fusion Model

We model the multimodal cyberbullying detection problem as follows. A multimodal cyberbullying detector uses n different data modalities $\{f_i, 1 \leq i \leq n\}$ (visual or textual features in our case) and outputs local decisions about cyberbullying incident C. In this work, we use the terms modalities and features interchangably. These decisions are represented by n probability values, $p_1, p_2, \ldots, p_n$, where $p_i = P(C|f_i)$ denotes the probability that cyberbullying event C has occurred based on modality f_i. A Bayesian approach is iteratively used to fuse these probabilistic decisions as follows. Let us consider that we have integrated a set $\mathbf{f}^{i-1}$ of $i-1$ modalities, i.e. f_1, f_2, $\ldots$, f_{i-1}, resulting into $P(C|\mathbf{f}^{i-1})$, which denotes the probability of occurrence of C based on a group of $i-1$ modalities. The individual decision based on i^{th} modality, i.e. $P(C|f_i)$, is integrated into $P(C|\mathbf{f}^{i-1})$ and the fused decision i.e. probability $P(C|\mathbf{f}^{i})$ of occurrence of C based on modality set $\mathbf{f}^{i}$ is calculated as [5]:

$$P(C|\mathbf{f}^{i}) = \frac{P^{+} \times \exp(\alpha_{f_i,\mathbf{f}^{i-1}})}{P^{+} \times \exp(\alpha_{f_i,\mathbf{f}^{i-1}}) + P^{-} \times \exp(-\alpha_{f_i,\mathbf{f}^{i-1}})} \quad (7)$$

where, P^{+} and P^{-} are the weighted combined probabilities of the occurrence and non-occurrence of cyberbullying, respectively, using $\mathbf{f}^{i-1}$ and f_i, and are given by:

$$P^{+} = P(C|\mathbf{f}^{i-1})^{\mathbf{w}_{i-1}} \times P(C|f_i)^{w_i} \quad (8)$$

$$P^{-} = (1 - P(C|\mathbf{f}^{i-1}))^{\mathbf{w}_{i-1}} \times (1 - P(C|f_i))^{w_i} \quad (9)$$

In the above equations, the exp term denotes the exponential function, and $\mathbf{w}_{i-1}$ and w_i are normalized confidence scores of $\mathbf{f}^{i-1}$ and f_i respectively and their sum is 1. The term $\alpha_{f_i,\mathbf{f}^{i-1}} \in [-1, 1]$ provides the degree of agreement/disagreement (called Agreement Coefficient [28]) between two modalities $\mathbf{f}^{i-1}$ and f_i, wherein -1 and 1 represent the full disagreement and the full agreement, respectively, between the two modalities. The computation of confidence scores and modeling of $\alpha_{f_i,\mathbf{f}^{i-1}}$ are described in the following paragraphs.

The proposed fusion model is adapted from [28] and it uses the logarithmic opinion pool (LOGP) consensus rule satisfying the assumption of conditional (content-wise) independence among different modalities [32]. The proposed model normalizes the outcome over the two aspects, the occurrence and non-occurrence, of a cyberbullying incident

(see denominator term in Eq. (7)). Underlying principle behind this fusion model is that the occurrence of a cyberbullying incident is determined with a higher overall probability when more concurring evidences are fused.

1) Accuracy of a modality: The confidence in a modality is related to how accurate it has been in the past. The higher the accuracy of a modality, higher the confidence we would have in it. Using the training data, we compute the accuracy (and therefore the confidence score) of a modality by determining how many times a cyberbullying incident is correctly detected based on it (using binary thresholding) out of the total number of incidents. Accuracy Acc_i for modality f_i is defined as follows:

$$Acc_i = P(|p_i - y_i| < 0.5) \tag{10}$$

2) Fairness of a modality: We define bias in a modality as the difference in the prediction accuracy between the privileged and the unprivileged groups i.e. $S = 1$ versus $S = 0$. We also use the shorthand notation of subscript $_{S=1}$ to denote the observations that correspond the privileged class and $_{S=0}$ for those in the unprivileged class.

$$Bias_i = Acc_{i,S=1} - Acc_{i,S=0} \tag{11}$$

In the above equation, $Bias_i$ denotes the bias for the i^{th} modality. Fairness is defined as the lack of bias.

3) Confidence score of a modality: Confidence score w_i for a modality is modeled as a combination of accuracy and bias.

$$w_i = Acc_i - \lambda \times Bias_i \tag{12}$$

where λ is a weighing parameter that captures the relative importance assigned to fairness and accuracy by the ML designer.

4) Combining confidence scores of multiple modalities: The overall confidence in a group of modalities is computed by combining the confidence scores of individual modalities. Let w_i and w_k be the confidence scores of two modalities f_i and f_k, respectively. A Bayesian method is used to compute the overall confidence w_{ik} in a group of these two modalities, as follows:

$$w_{ik} = \frac{w_i \times w_k}{w_i \times w_k + (1 - w_i) \times (1 - w_k)} \tag{13}$$

The above formulation is based on assumption that although the modalities are correlated in their decisions; they are mutually independent in terms of their confidence scores [28].

Eq. (13) can be extended for n modalities [5], by replacing the confidence scores of two modalities (i.e. w_i and w_k) with that of two groups of modalities, say $\mathbf{w}_{i-1}$ representing the overall confidence scores of a group of $i-1$ modalities and w_i denoting the confidence score of i^{th} modality.

5) Agreement coefficient between modalities: The agreement coefficient $\alpha_{i,k}$, between the modalities f_i and f_k is modeled based on Pearson's moment correlation, as follows:

$$\alpha_{i,k} = \frac{\sum_{j=1}^{n}(p_{i,j} - \bar{p_i})(p_{k,j} - \bar{p_k})}{\sqrt{\sum_{j=1}^{n}(p_{i,j} - \bar{p_i})^2}\sqrt{\sum_{j=1}^{n}(p_{k,j} - \bar{p_k})^2}} \tag{14}$$

where, p_i and p_k are the individual probabilities of the occurrence of cyberbullying based on modalities f_i and f_k, respectively. These probabilities represent decisions about the detection tasks. The full agreement ($\alpha_{i,k} = 1$) occurs when there are exactly same probabilities. On the other hand, the two modalities are considered to be in full contradiction with each other ($\alpha_{i,k} = -1$) when there exists totally dissimilar probabilities.

To determine the agreement coefficient between the two sets $\mathbf{f}^{i-1}$ and f_i of modalities, we adopt the concept of *average-link clustering* [33], which considers the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. In our case, we have two clusters: a group $\mathbf{f}^{i-1}$ of $i-1$ modalities and a new i^{th} modality. The following equation is used for agreement fusion:

$$\alpha_{f_i,\mathbf{f}^{i-1}} = \frac{1}{i-1}\sum_{s=1}^{i-1}\alpha_{s,i} \tag{15}$$

Here, $\alpha_{s,i}$ for $1 \leq s \leq i-1$, $1 < i \leq n$ are the agreement coefficients between the s^{th} and i^{th} modalities.

V. Implementation Details and Dataset

We validate our method using an Instagram data set (image, text captions, and text comments from others) made available by Hosseinmardi et al. [12] that has been used in different studies of cyberbullying [13], [20]. This dataset contains about 2000 media sessions from Instagram, including the posted image, caption, and comments from other users. Hosseinmardi et al. [12] used a snowball sampling method to identify Instagram ids. Each image and the accompanying comments were considered a "media session". Only media sessions which contained more than 15 comments were considered eligible for inclusion in this dataset. Each media session in the resulting dataset was hand-labelled by five crowd-sourced (CrowdFlower) annotators for the presence of cyberbullying.

However, we found only 699 of these media sessions to still have images accessible from the recorded Instagram URLs. These 699 labeled media sessions were used in [13]. (Note that [13] did not study the problem of fairness nor employ a Bayesian fusion framework.) As is often the case in cyberbullying datasets, this dataset was skewed and has only 119 bullying instances as opposed to 580 non-bullying instances. To maintain focus on the fairness aspect in this work, we consider a balanced subset of data (N=238)

Table I: Difference in the performance for the privileged and under privileged groups in the baseline methods

Approach	ΔAUC (%)	AUC T-test p-value	ΔTPR (%)	TPR T-test p-value	ΔFPR (%)	FPR T-test p-value
Naïve Bayes	6.45	< 0.001	4.86	< 0.001	1.32	0.54 (n.s.)
Our Baseline ($\lambda = 0$)	8.65	< 0.001	6.21	< 0.001	11.08	< 0.001

which contains equal number of bullying and non-bullying instances.

The dataset includes the textual and visual features as derived in [13]. This includes textual features indicative of emotion, gender specific terminology, sexual connotations, as well as the relative distribution of different parts of speech. Specifically, the authors used Linguistic Inquiry and Word Count (LIWC) to analyze the text for cyberbullying. LIWC provides more than 90 descriptive variables which include word counts and language use characteristics [34]. The list of features includes psychological process, tone, word count, presence of informal language, use of third person pronouns, use of sexual words, and use of words indicating violence.

The authors also used Microsoft's Project Oxford to extract visual features from the Instagram images [35]. Project Oxford is a computer vision API which analyzes the images for the dominant colors, number of people present, category, adult content, etc. The features computed included age and gender of people present in the image, image category (e.g., presence of tattoos, graffiti, drugs, generic image labels) and image type (e.g., colored, black and white, clipart). The original dataset consisted of 204 textual and visual features. For "Our Method" and "Our Baseline" implementations in this work, a feature selection criteria was applied based on feature weight > 0.60 as computed using Eq. 12. For "Our Method" this resulted in a subset of 34 features.

One of the features in the dataset was the apparent gender of the person in the image as identified by the computer vision API. As there is significant literature suggesting that women tend to be marginalized in cyberbullying instances [36], [37], we consider gender of the person in the image to be a sensitive attribute and male to be the privileged class. Non-male identifications are considered the unprivileged group. In the considered dataset, there are 60 instances of (only) male subjects and 188 instances of non-male subjects. Note that the non-male subjects category includes instances where gender was not clearly identified by the API or more than one gender was present in the image (e.g., group images).

VI. Results

We implemented the approach proposed in Section IV using Python programming language and tested it on the dataset described in Section V. The continuous features were discretized based on deciles before analysis. The training:test split was done in the ratio of 75:25 and 100 iterations with different train-test splits were undertaken and the results averaged over those iterations.

We also implemented two baseline approaches to compare with the performance of the proposed approach. This includes: (a) Naive Bayes, which is also based on a Bayesian approach and is a well-known machine learning method [38]; and (b) "Our Baseline" which includes the approach described in Section IV but with the value of λ (see Eq. 12) set to 0. Hence, while it makes uses of the association between features and the confidence scores for better fusion, it gives zero weight to fairness of the features.

A. Auditing Current Approaches (Baselines) for Bias

Consistent with the literature on fairness in machine learning we use three primary metrics: AUC (area under the receiver operating curve; a robust alternative to prediction accuracy [39]), TPR, and FPR in this work. While their average value across the two groups (privileged and unprivileged) is used to quantify accuracy, the deltas (Δ) between the two groups are used to quantify bias. An ideal approach will yield a very high accuracy (e.g, overall AUC) and very low bias (e.g., ΔAUC).

We found significant differences in the performance of the baseline algorithms for the privileged and unprivileged groups. For instance, the percentage difference for ΔAUC was 6.45% for the Naive Bayes approach and 8.65% for "Our Baseline". We also conducted pairwise T-test to check if these differences are statistically significant. The results, i.e. the differences across the groups for AUC, TPR, and FPR are summarized in Table I.

As can be seen in Table I, the performance was noticeably different in terms of almost all the metrics for both the baseline approaches. The only exception was *not significant* (n.s.) t-test result for ΔFPR (1.32%) for Naive Bayes. However, the observed FPR scores for both privileged and unprivileged group were quite high ($> 35\%$, see Table II) in that case and hence this result is likely unacceptable in practical settings. While the FPR was lower for "Our Baseline" approach it had a significant Δ of 11.08%. As such, both the baselines seem to have significant bias and/or accuracy issues, thus motivating the need for a newer approach that can improve fairness and accuracy.

B. Impact of the Bias Reduction Approach

Next, we implemented the proposed approach with the goal of reducing the deltas between privileged and unprivileged groups while keeping the accuracy levels high. The value of λ in the proposed approach was selected based empirical testing. The value was varied in the range 0.01 to 10 in increments of 0.01, and the value yielding the highest

Table II: Performance of two baselines and the proposed approach

	AUC Privileged	AUC Unprivileged	TPR Privileged	TPR Unprivileged	FPR Privileged	FPR Unprivileged
Naive Bayes	84.86	78.41	85.61	80.75	36.31	37.63
Our Baseline ($\lambda = 0$)	88.74	80.08	87.88	81.67	10.41	21.49
Our Method ($\lambda = 3.89$)	83.80	83.32	83.60	83.95	16.01	17.32

Table III: Differences in the performance for privileged and unprivileged groups

Approach	ΔAUC (%)	AUC T-test p-value	ΔTPR (%)	TPR T-test p-value	ΔFPR (%)	FPR T-test p-value
Our Baseline ($\lambda = 0$)	8.65	< 0.001	6.21	< 0.001	11.08	< 0.001
Our Method ($\lambda = 3.89$)	0.48	0.18 (n.s.)	0.36	0.35 (n.s.)	1.31	0.11 (n.s.)

accuracy was selected. It was found to be 3.89 in the current dataset.

This approach resulted in better accuracy AUC, TPR, and FPR values for the proposed approach compared to both the baselines (see Fig. 1). It also resulted in better fairness scores (reduced Δ in the performance for the privileged and unprivileged group) across different metrics of ΔAUC, ΔTPR, ΔFPR as can be seen in Fig. 2. The detailed scores for the privileged and unprivileged groups are available in Table II and the Δ values are summarized in Table III.

To consider one example, the Δ of the AUC score between the privileged and unprivileged groups in the proposed approach is 0.48%, which is more than 18 times lower than "Our Baseline" approach (8.65%) and more than 13 times lower than the Naive Bayes approach (6.45%). We see similar trends in terms of TPR, where our method yields a ΔTPR of 0.36% which is multiple times lower than Naive Bayes and "Our Baseline". For FPR, "Our Method" yields a Δ of 1.31%, which is multiple times smaller than "Our Baseline" but is only marginally smaller than the Naive Bayes approach. However, as mentioned earlier, the high false positive rate for Naive Bayes ($> 35\%$) makes it unsuitable for a practical application (see Fig. 1).

Hence, the trends indicate that the proposed approach was able to make noticeable reductions in the Δs across privileged and unprivileged groups while keeping the accuracy levels (in terms of AUC, TPR, FPR) high. The fact that the accuracy performance did not drop in terms of any of the metrics indicates that the goals of fairness and accuracy need not be orthogonal to each other, and the feature selection process may have been able to weed out some of the less useful features based on the stricter (accuracy + fairness based) thresholds.

Lastly, we test if the proposed approach is able to reduce the statistically significant differences across the two groups as were observed in the baseline condition and discussed in Section VI-A. We undertook another round of difference of means pairwise T-test on the outputs of the proposed approach. The results are summarized in Table III (with the results for "Our Baseline" presented to allow for comparison). The p-values for all the difference of means test were *not* significant in "Our Method" even though they were significant in the baseline condition. This indicates that the proposed approach has been able to render the group differences insignificant, which was a primary goal of the proposed approach.

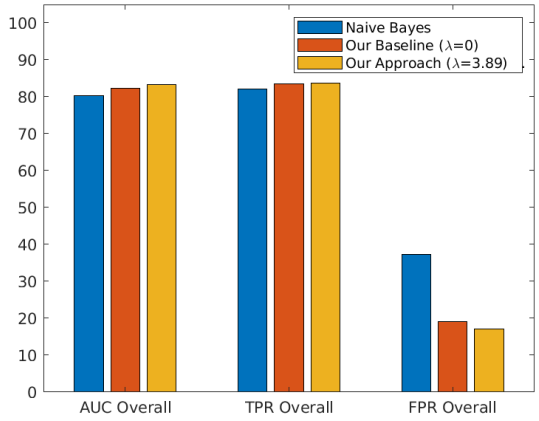

Figure 1: Overall performance of different approaches (Baselines + Proposed)

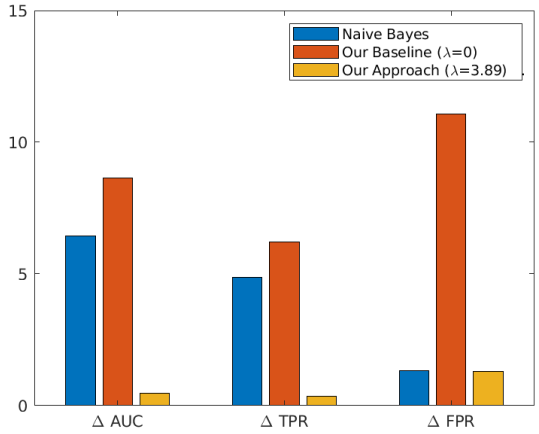

Figure 2: Δ of performance for privileged and unprivileged groups across different models

Based on the trends in the datasets, we report that the proposed weighted probabilistic fusion-based approach is

beneficial at decreasing the disparity in the performance of fusion algorithm across gender as computed through the metrics of ΔAUC, ΔTPR, and ΔFPR. Further, we see the average accuracy level increased with the proposed weighted probability fusion approach.

This work also has a number of limitations. First, this work uses a relatively small, balanced dataset from a single social media platform. Further, the work uses a narrow definition of privileged class – one that is based on inferred gender of the person in the image. However, note that proposed fairness aware fusion approach is generic and can easily be applied to other datasets and other definitions of sensitive attributes. Future work should include non-binary identities as well as other notions of demography such as race, age, nationality, etc. in a similar fairness analysis.

VII. Conclusion

This paper describes one of the first attempts at a Bayesian fusion framework that not only optimizes for accuracy but also considers fairness. The framework takes into account the accuracy and the fairness score for each modality to assign them weights. The weights of each modality and the agreement between them is used to come up with optimal decisions that balance accuracy and fairness. The results of applying the framework to a multimodal (visual + textual) cyberbullying detection problem demonstrate the efficacy of the approach in yielding high levels of both accuracy and bias. The results pave way for a more accurate and fair approach for cyberbullying detection, which would provide equitable opportunities to different groups in improving their quality of life.

Acknowledgment

The work of Jamal Alasadi was supported by the MOHESR, Iraq. Work by Ramanathan Arunachalam and Vivek Singh was supported in part by the US National Science Foundation under Grant SES-1915790.

References

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[2] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–26, 2018.

[3] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *fifth international AAAI conference on weblogs and social media*, 2011.

[4] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine learning and applications and workshops*, vol. 2. IEEE, 2011, pp. 241–244.

[5] V. K. Singh, Q. Huang, and P. K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016, pp. 884–887.

[6] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, 2014, pp. 3–6.

[7] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in vine," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2015, pp. 617–622.

[8] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "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*, 2015, pp. 280–285.

[9] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, 2018, pp. 77–91.

[10] E. Raisi and B. Huang, "Reduced-bias co-trained ensembles for weakly supervised cyberbullying detection," in *International Conference on Computational Data and Social Networks*. Springer, 2019, pp. 293–306.

[11] V. K. Singh and C. Hofenbitzer, "Fairness across network positions in cyberbullying detection algorithms," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2019, pp. 557–559.

[12] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," *arXiv preprint arXiv:1503.03909*, 2015.

[13] V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 2090–2099.

[14] V. K. Singh, M. Chayko, R. Inamdar, and D. Floegel, "Female librarians and male computer programmers? gender bias in occupational images on digital media platforms," *Journal of the Association for Information Science and Technology*, 2020.

[15] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 3, pp. 1–30, 2012.

[16] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 339–347.

[17] Hitkul, R. R. Shah, P. Kumaraguru, and S. Satoh, "Maybe look closer? detecting trolling prone images on instagram." in *BigMM*, 2019, pp. 448–456.

[18] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models; a reproducibility study," *arXiv preprint arXiv:1812.08046*, 2018.

[19] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, 2017.

[20] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in *International conference on social informatics*. Springer, 2015, pp. 49–66.

[21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[22] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[23] J. Alasadi, A. Al Hilli, and V. K. Singh, "Toward fairness in face matching algorithms," in *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, 2019, pp. 19–25.

[24] Y. Bechavod and K. Ligett, "Penalizing unfairness in binary classification," *arXiv preprint arXiv:1707.00044*, 2017.

[25] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 289–295.

[26] Y. Wang, C. Von Der Weth, Y. Zhang, K. H. Low, V. K. Singh, and M. Kankanhalli, "Concept based hybrid fusion of multimodal event signals," in *2016 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2016, pp. 14–19.

[27] F. Kong, Y. Zhou, and G. Chen, "Multimedia data fusion method based on wireless sensor network in intelligent transportation system," *Multimedia Tools and Applications*, pp. 1–13, 2019.

[28] P. K. Atrey, M. S. Kankanhalli, and R. Jain, "Information assimilation framework for event detection in multimedia surveillance systems," *Multimedia systems*, vol. 12, no. 3, pp. 239–253, 2006.

[29] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki, "Fair adversarial gradient tree boosting," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 1060–1065.

[30] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[31] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[32] C. Genest, J. V. Zidek *et al.*, "Combining probability distributions: A critique and an annotated bibliography," *Statistical Science*, vol. 1, no. 1, pp. 114–135, 1986.

[33] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[34] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[35] "Microsoft Project Oxford," https://blogs.microsoft.com/ai/microsofts-project-oxford-helps-developers-build-more-intelligent-apps/.

[36] J. F. Chisholm, "Cyberspace violence against girls and adolescent females," *Annals of the New York Academy of Sciences*, vol. 1087, no. 1, pp. 74–89, 2006.

[37] P. K. Smith, F. Thompson, and J. Davidson, "Cyber safety for adolescent girls: bullying, harassment, sexting, pornography, and solicitation," *Current opinion in obstetrics and gynecology*, vol. 26, no. 5, pp. 360–365, 2014.

[38] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[39] C. X. Ling, J. Huang, H. Zhang *et al.*, "Auc: a statistically consistent and more discriminating measure than accuracy," in *Ijcai*, vol. 3, 2003, pp. 519–524.