Mitigating demographic bias of machine learning models on social media

Yanchen Wang
yw516@georgetown.edu
Georgetown University
Washington, DC, USA

Lisa Singh
Lisa.Singh@georgetown.edu
Georgetown University
Washington, DC, USA

ABSTRACT

Social media posts have been used to predict different user behaviors and attitudes, including mental health condition, political affiliation, and vaccine hesitancy. Unfortunately, while social media platforms make APIs available for collecting user data, they also make it challenging to collect well structured demographic features about individuals who post on their platforms. This makes it difficult for researchers to assess the fairness of models they develop using these data. Researchers have begun considering approaches for determining fairness of machine learning models built using social media data. In this paper, we consider both the case when the sensitive demographic feature is available to the researcher and when it is not. After framing our specific problem and discussing the challenges, we focus on the scenario when the training data does not explicitly contain a sensitive demographic feature, but instead contains a hidden sensitive feature that can be approximated using a sensitive feature proxy. In this case, we propose an approach for determining whether a sensitive feature proxy exists in the training data and apply a fixing method to reduce the correlation between the sensitive feature proxy and the sensitive feature. To demonstrate our approach, we present two case studies using micro-linked Twitter/X data and show biases resulting from sensitive feature proxies that are present in the training data and are highly correlated to hidden sensitive features. We then show that a standard fixing approach can effectively reduce bias even if the sensitive attribute needs to be inferred by the researcher using existing reliable inference models. This is an important step toward understanding approaches for improving fairness on social media.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → *User characteristics.*

KEYWORDS

algorithmic fairness, social media, demographic inference

ACM Reference Format:
Yanchen Wang and Lisa Singh. 2023. Mitigating demographic bias of machine learning models on social media. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30–November 01, 2023, Boston, MA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3617694.3623244

EAAMO '23, October 30–November 01, 2023, Boston, MA, USA

ACM ISBN 979-8-4007-0381-2/23/10.
https://doi.org/10.1145/3617694.3623244

1 INTRODUCTION

Most fairness studies use well structured data sets, where both the attributes and the properties of the individuals, e.g. demographic features, are clearly specified and understood. While an important first step, we know that computer scientists and data scientists are using machine learning algorithms in considerably more complex contexts. In these complex contexts, there may be missing values, poor quality attributes, or a wider range of data types, including text and images. Just as critical, there may also be missing information about sensitive features. In fact, in some contexts, the sensitive feature of interest may never be available to researchers, making it difficult to determine the fairness of a model. In this paper, we consider one such context - social media.

Researchers have been using social media data to better understand different behaviors and opinions of individuals, including extremism [1, 59], mental health conditions [12, 30], and political affiliation [16, 35]. Many of these types of studies make use of opportunistic data collection from APIs. Unfortunately, while social media platforms make APIs available for collecting data (including posts, user account information, and network information) from their platforms, they also make it challenging to collect well structured demographic features about the individuals using their platforms. These features are needed by researchers to assess the fairness of models built using these data.

To further exacerbate the situation, platforms maintain different demographic characteristics about their users. Singh and colleagues studied required demographic information across some popular social media platforms, including Twitter/X, Facebook, Tiktok, LinkedIn and Snapchat, and found large variability in the type and amount of demographic information required to create an account on these platforms [56]. For example, gender and birth date are required to open an account on Facebook, but neither are required when creating an account on LinkedIn (location is required on LinkedIn). This inconsistency in the requirements of demographic data means that even if demographic data are made available by platforms, much of the demographic data would contain missing values, limiting researchers ability to access the demographic characteristics they need in order to measure potential bias in their training data, and ultimately in their learning models.

The sharing of demographic characteristics leads to an additional issue, user privacy. While there are privacy concerns associated with having access to some sensitive demographic attributes, there are potential fairness concerns associated with not having access, or having inconsistent access, i.e. platforms on which some users

share their demographic information, but others do not. To better understand the impact of these issues on fairness, this paper investigates the fairness of machine learning models constructed using social media data. We consider scenarios when missing data exist because some users have accounts but do not actively post, and the sensitive attribute of interest is not readily available to the researcher and must be inferred. Our work compliments existing studies since we conduct our empirical evaluation on micro-linked social media data. Specifically, we use Twitter/X data for which we have demographic data through a micro-linked survey. Users in our study have given consent to micro-link their survey response to their Twitter/X data (see Section 7 for more detail).

More specifically, after describing challenges associated with measuring fairness on social media, we demonstrate the existence of fairness issues using two Twitter/X case studies.[1] Consistent with previous work, we find that even if sensitive features are not used to build the machine learning model, bias exists because non-sensitive features serve as proxies to these sensitive features. We refer to these non-sensitive features as *sensitive feature proxies*. We then consider an approach for reducing these biases when the ground truth sensitive attribute is available and when it must be inferred. Our empirical evaluation shows that our approach can effectively reduce bias even if the sensitive attribute of interest is not initially available to the researcher.

Our main contributions can be summarized as follows. (1) We formally define and describe hidden sensitive features and sensitive feature proxies for social media data and discuss the additional challenges social media data pose when trying to determine model fairness. (2) We introduce an approach to measure the relationship (overlap) between the sensitive feature proxy and the hidden sensitive feature. (3) Using two Twitter/X case studies, we show the bias that manifests within classification models as a result of this relationship, including bias from missing data, i.e. low activity level. (4) We consider an existing fixing approach, apply it to this new social media setting, and show that it improves fairness with a very small tradeoff in performance, even when the sensitive attribute is not readily available to the researcher.

2 RELATED LITERATURE

Numerous examples of discriminatory or unfair machine learning models and applications have been described in the literature [19, 40, 41]. In the USA, the Civil Rights Act of 1964 states that it is illegal to discriminate against people based on race, color, religion, sex, or national origin. These demographic traits are examples of sensitive/protected attributes or attributes that should not be dominant features used by machine learning algorithms to make predictions. Therefore, in order to avoid discrimination, information about sensitive attributes must be available, and researchers must use this information to quantify bias in classifiers. We begin this section by describing ways researchers measure fairness and make adjustments when models are not fair. We then present literature related to social media fairness. Traditional fairness research has focused on numeric and/or categorical features, but social media data sets are typically textual. Therefore, we focus our discussion on fairness in text.

2.1 Measuring and improving fairness

In machine learning fairness, there are two main families of fairness definitions: group-based and individual-based [15]. The group-based or statistical notion of fairness use demographic parity, equalized odds, and CV to measure fairness [10, 24]. The individual notion of fairness attempts to ensure that similar individuals are treated similarly [22]. There is no standard way to measure this as it is very context dependent. In this paper, we consider group-based fairness and use group-based fairness metrics to measure bias in the machine learning models we consider.

To correct bias, a number of correcting algorithms have been proposed. Correction algorithms can be grouped into three categories: pre-processing, in-processing and post-processing. Pre-processing approaches attempt to fix the input data in order to improve fairness [9, 11, 24, 37, 45, 58]. For example, Feldman et al. fix continuous data by changing feature values to remove correlations between feature values and the sensitive attribute [24]. Wang and Singh focus on the role of missing values and selection bias on the fairness of categorical attributes and use resampling and reweighting to improve fairness [58]. In-processing fixes add fairness constraints in the objective function during the training process [5, 38, 47, 57, 63]. For example, researchers have introduced changes to some traditional machine learning models, such as logistic regression and decision trees, to include a fairness constraint in the objective function to ensure that each subgroup has an equal probability of receiving a positive outcome [38]. Post-processing approaches change predicted labels after the model is trained. They modify the results of a trained classifier to ensure fair prediction results on sensitive attributes [8, 29, 31, 50, 51]. For example, Petersen et al. use the predicted results and a similarity graph between individuals to ensure individual fairness, i.e. similar people should be treated similarly and get similar predicted results [50].

Some bias correction techniques have been applied to text classification. Geyik et al. quantify gender bias in the LinkedIn search ranking algorithm and propose a mitigation algorithm that changes the existing ranking algorithm to ensure fairness based on a user's sensitive attribute. If one demographic group has a much lower rank than others, the algorithm gives that group a higher rank to ensure fairness [26]. Bolukbasi et al. find that word embeddings trained on Google News articles exhibit gender bias and they propose a fixing algorithm that first identifies words with gender bias and then adjusts word vectors to remove the bias [8].

While most popular fairness toolkits (such as AI Fairness 360 by IBM [4], Fairlearn by Microsoft [7], and the What-if tool by Google [60]) require ground truth information about the sensitive attribute, there are bias correction algorithms that do not require sensitive attribute information. One approach is to use proxies for sensitive attributes based on other features in the training data [13, 17, 18, 53, 64]. For example, Zhang [64] proposes using the geo-location and the last name of applicants to infer their race and then uses the inferred race to ensure that a race bias does not exist during the mortgage application process. Using proxies for sensitive attribute could still cause fairness issue. For example, using the geo-location

[1]We focus on Twitter/X since the platform is used by many researchers and the data being analyzed are all public.

and the last name to infer race could be biased against women who marry outside their ethnic group [48, 61]. Chen et al. propose using proxy features to measure fairness without using the ground truth sensitive attribute [13]. Another approach is to use privacy-preserving methods to avoid directly using sensitive attributes [32, 62]. For example, Hu and colleagues [32] propose a distributed, privacy-preserving fair learning framework using multiple local agents, where each local agent separately holds different sensitive demographic data. During the training process each agent learns a fair local dictionary and sends it to the modeler. The modeler then learns a fair model based on an aggregated dictionary. While both of these strategies are promising, our approach differs since we focus more on improving the fairness and we assume a domain that contains noisy, partial (missing) data. Furthermore, in our data, we do not have a separate proxy that we know about apriori to use to approximate the sensitive attribute. Instead our study assumes that the training data includes a hidden sensitive attribute proxy that we need to adjust for if we want to build a fair classifier.

2.2 Machine learning model fairness using social media data

Different researchers have identified fairness and ethical issues arising from models built using social media data [42, 56]. Ekstrand et al. use accuracy differences across demographic groups as a fairness metric and find that researchers often fail to check if every subgroup is treated fairly while evaluating the effectiveness of recommendation systems. The authors show that female users receive less accurate results than male users [23]. Sherman et al. use a population of Reddit users with known genders and depression statuses to analyze gender bias in depression classifiers and discuss potential correction methods, e.g. adding a regularization term for gender to correct for the performance differentials across demographic groups [55]. Ball-Burack et al. show racial dialect bias in harmful tweet detection. Specifically, the African American English dialect tweets have a higher rate of being detected as harmful compared to White English tweets. They use existing bias mitigation methods such as debiasing word embedding and reweighting to correct for the racial bias in the harmful tweet detection algorithm [2].

We pause to mention that while some of the mentioned work proposes fixing approaches, most fixing methods have not been designed using social media data. For example, the gender neutral word embeddings introduced by Bolukbasi et al. are built using text from newspapers [8]. Social media posts are different from newspaper articles since social media language is less formal and posts are shorter than articles. Previous research has shown that embeddings built on formal text often do not have good performance on social media data [39], reminding us that much work still exists to translate general text fixing approaches to ones that are effective for models constructed using social media data.

Researchers need to be able to measure fairness on social media even when sensitive features are not available. To support this scenario, we highlight some research that infers demographic information from social media with high accuracy [14, 43, 44]. Many researchers use classic machine learning algorithms for demographic inference. Newer work uses neural models for demographic inference. For example, Liu et al. predict gender and age from user posts and biography data using BERT (Bidirectional Encoder Representations from Transformers) to transform text into embeddings and use the embeddings as input to build a neural network with LSTM (Long-short term memory) and attention layers [44]. We will use this approach to infer the sensitive feature of interest and demonstrate that proxies exist for these features in the training data set (see Section 4).

Ghosh et al. investigate how uncertainty and errors in demographic inference impact the fairness in fair ranking algorithms. They use race and gender as sensitive attributes and suggest that researchers should not use inferred demographic data as input unless the inference results have very high accuracy [27]. We acknowledge that high accuracy is necessary to use inferred attributes. This is one reason we focus on gender. It is a demographic that can be inferred with very high accuracy.

3 SOCIAL MEDIA CLASSIFICATION

For many social media classification tasks different forms of text are used as training data. For example, we may use posts to predict a person's likelihood of getting vaccinated. However, these posts may be highly correlated to sensitive demographic features like gender or race, and may lead the classifier to perform differently across different race and gender groups. When we do not have access to these sensitive demographic features, we will refer to them as *hidden sensitive features* within our learning environment. In this section, we describe the relationship between our learning task and these hidden sensitive features (Section 3.1). We then present the challenges that arise for social media classification as a result of the hidden sensitive features and data issues that could negatively affect fairness when building models with social media data (Section 3.2).

3.1 Modeling hidden sensitive features

Figure 1a shows a diagram of how a classic machine learning model is generated. In this work, our prediction task is binary. Let $Y = \{y_1, y_2, \ldots, y_n\}$ be the set of binary labels we want to predict and for the ith observation, $y_i \in \{+, -\}$ with $y_i = +$ being a positive or desired outcome and $y_i = -$ being a negative or non-desired outcome. Similarly, let $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ be the set of predicted label and $\hat{y}_i \in \{+, -\}$ represent the predicted label of ith observation.[2] Given a labeled training set consisting of training features X and labels Y, a machine learning model M is constructed. We use the features in the training data X (blue circle on the left) to predict labels $\hat{Y}$ (blue rectangle on the right) using machine learning model M, where $\hat{Y} = M(X)$, and evaluate the differences between Y and $\hat{Y}$.

Previous literature has shown that language usage is correlated with different demographics, including sensitive ones like gender and race [54]. Because we will use text for our prediction task, we want to model this possible relationship between a sensitive demographic feature and features constructed from text. Let S = $\{s_1, s_2, \ldots, s_n\}$ be the set of values for a binary sensitive feature, e.g. gender, that is known to the researcher, but not part of the training data. The set S contains n individual observations, and

[2]We note that for some inference tasks on social media, there is not a clear privileged outcome. In those cases, we typically define the minority class (in terms of available labels) as the privileged group.

s_i represents the sensitive feature value of the ith observation.[3] For the ith observation, we set $s_i = 1$ if the observation is in the privileged group and $s_i = 0$ if it is in the unprivileged group. Figure 1b shows the training and prediction process with this hidden sensitive feature, where S is shown in the triangle. Similar to Figure 1a, the big blue circle represents the training features. In this figure, we also show a smaller circle within the blue circle that represents the subset of training features (text features in our case) X^p that are highly correlated with the hidden sensitive feature S. We refer to X^p as a *sensitive feature proxy*. When this situation exists, we say that a strong relationship exists between X and S through X^p. In other words, while S is not in X, proxies for it are and these proxies are being used to build $M(X)$ and predict Y, thereby possibly biasing the inference process.

To build a fair classifier, we want to minimize the size of the subset of features in X that are highly correlated to sensitive attribute S. This is illustrated in Figure 1c, where the size of X^p is smaller and the correlation between X^p and S is insignificant. This means that the bias generated by using a proxy of S has been reduced, model $M'(X)$ will not use "hidden" sensitive information or a sensitive proxy to predict $\hat{Y}$, and therefore, $M'(X)$ will be more fair than $M(X)$.

3.2 Challenges associated with social media classification fairness

If we know the demographics of the users in X, then we can quantify the bias in M and determine an appropriate fixing algorithm to ensure that M' is less biased. Unfortunately, as mentioned in Section 1, demographic data is not consistently available on different platforms. Platform users also share different amounts of demographic information, meaning that while researchers may have access to the demographic characteristics of some people within their sample, it is highly likely that a large fraction of their sample share different demographic characteristics, leading to a large fraction of missing values.

Researchers have used different approaches for obtaining these demographics uniformly across their data sets. One approach is to obtain demographic information through surveys [49] and obtain informed consent to link the survey information to their Twitter/X account. However, for analyses that involve millions of platform users, this approach is prohibitively expensive, especially given low survey response rates [1]. Another approach is to use the shared features to infer the demographic feature of interest. We will refer to the inferred value of S as $\hat{S}$. Using this approach leads to additional secondary questions. Can we infer the hidden sensitive feature S and use that information to accurately measure the bias created by the sensitive feature proxy X^p that informs the researcher about whether or not $M(X)$ is fair? If the answer is yes and the model is not fair, can we make adjustments to ensure that $M'(X)$ is fair and accurate?

[3]If the sensitive feature is not binary, it can be converted into a binary feature by defining a *privileged* group and an *unprivileged* group. For example, assume the sensitive feature S is race and that there are five values for race in the data set: {`White`, `Black`, `Hispanic`, `Asian`, `Others`}. We can convert these values into a binary attribute where one or more races are designated as the privileged group, and the other races as the unprivileged group. There is no universal rule for which attributes go into the privileged group - it depends on the sample and the learning task.

An additional issue that arises with respect to inference using social media data results from different levels of user activity on the platform. Users spend varying amounts of time using social media and post at different rates. Some users choose to actively share their opinions on social media, while others rarely post. Previous research has shown that on Twitter/X, the most active 10% of users who are U.S. adults contribute 80% of all tweets created [33]. It has also been shown that more information can be inferred from active users than relatively inactive ones [46]. We suggest using activity level as a new way to think about missing data. If the activity level of a user is too low, meaning the amount of missing data is particularly high, then reliable inference is not possible using that user's posts. Finally, selection bias is another data issue that is prevalent on social media since different people choose to join different social media sites. Selection bias happens if observations from some groups in the sample are oversampled and others are undersampled. In our case studies, we consider the impact of each of the challenges.

4 METHODOLOGY

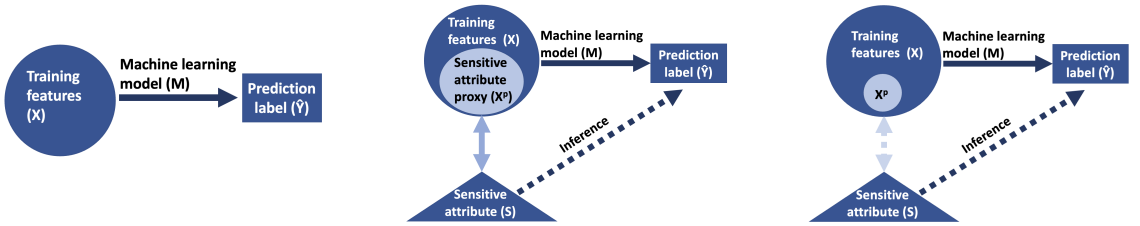
In this section, we present our approach, focusing on the experimental design (Section 4.1), and the models we consider in our empirical evaluation (Section 4.2).

4.1 Experimental design

We present our high level experimental design in Figure 2. We begin by training each machine learning model using text and numerical features (step A). We then compute the performance of each model (accuracy and F1 score), and the fairness of the model (p%-rule and CV) using S (step B). We note that S is not contained in X. We only use it to determine the fairness of $\hat{Y}$. We then identify the model that has the highest accuracy and F1 score (step C). If the fairness of this model is low, we say that a strong correlation exists between X^p and S. When this occurs, we apply the pre-processing fix method proposed by Wang and Singh [58] (step D). This method removes selection bias in the training data by randomly removing examples from the over-represented group and adding examples using resampling to the underrepresented group. This fixing methods will increase statistical independence between the sensitive feature S and the outcome Y. We then retrain the best model M(X) using the new data set that contains less selection bias and use fairness metrics to evaluate whether or not the model is more fair (step E).

As mentioned in Section 3, sensitive demographic information is often not available in social media data sets. In this second scenario, we select an existing demographic inference model (step F), and then determine $\hat{S}$ by inferring the demographic feature of interest for each user (Step G). We then assume $\hat{S}$ is the ground truth, and apply the same fixing method to remove selection bias.

The importance of this second experiment is twofold. First, we can simulate how fairness can be determined when the demographic information is not available for researchers to use. Second, because we have access to the actual ground truth labels in our data set, we can compare the difference in fairness quality of the prediction models using the actual demographic information and the inferred demographic information.



(a) Typical machine learning model

(b) Machine learning model with inferred sensitive attribute

(c) Machine learning model with inferred sensitive attribute

Figure 1: Machine learning model training

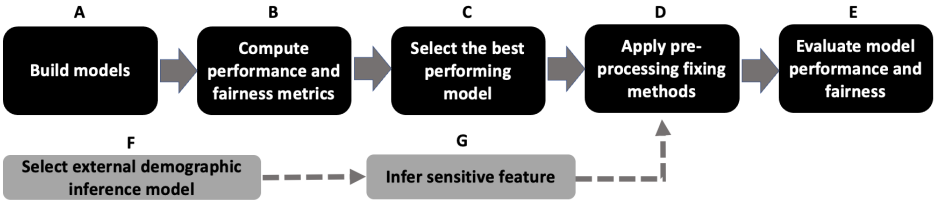


Figure 2: High level experimental design

Table 1: Training features

Category	Features
Account statistics	Number of tweets, number of days since first tweet
Network statistics	Number of followers, number of followings
Tweet statistics	Average number of words per tweet, average word length per tweet, proportion of emojis per tweet, proportion of hashtags per tweet, proportion of punctuation per tweet, proportion of emojis in biography, proportion of hashtags in biography, proportion of punctuation in biography
Biography statistics	Number of words in biography, average word length in biography proportion of emojis in biography, proportion of hashtags in biography, proportion of punctuation in biography
N-gram features	Unigram, bigram, trigram

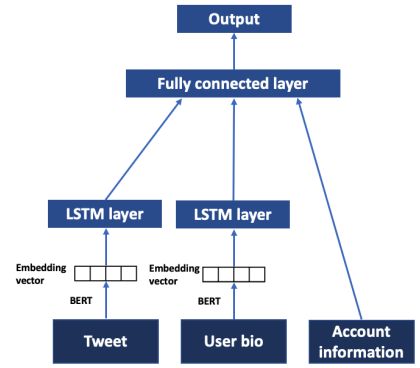
4.2 Feature Construction and Machine Learning Models

For our analysis, we use two classes of machine learning algorithms, classic machine learning models with N-gram features and neural network models with BERT embeddings.

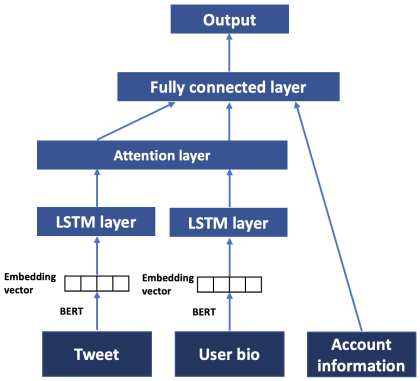
Classic machine learning model with N-gram features. We use user account statistics, user network statistics, user biography, and user tweets to build our machine learning models (see Table 1). For the user tweets and biography, we extract numerical features such as number of emojis in each tweet and create text tokens using N-gram (unigrams, bigrams and trigrams). We then use the token frequency as features to train our models. We train classic machine learning models that have performed well on different prediction tasks using Twitter/X data: logistic regression, random forest, decision tree, and support vector machine (SVM). While not exhaustive, this is a representative set.

Neural network model. In addition to classic machine learning models, we want to explore fairness of deep learning models. Figure 3 shows the architecture of the neural network models we use, a standard BERT model and a BERT model with an attention layer. Both models use BERT, a pretrained contextual language representation that can be used to convert text into embeddings [20]. Researchers have shown that BERT has good performance on traditional NLP tasks, including demographic inference on Twitter/X [44]. For the standard BERT model, we convert tweets and user biographies into embeddings using the pretrained uncased BERT-Base model [52]. We then input embedding vectors into a Long Short Term Memory (LSTM) layer and after the LSTM layer, we add account information features together into a fully connected Multi-Layer Perceptron neural network with one hidden layer.

The second model is an attention based model. In text classification tasks, some types of information may be more important for the learning task than other information. For example, if we are predicting a user's opinion or stance on an issue, posts may be more important than a user's account information. In an attention mechanism neural network model, models learn what subsets of the information are more informative and use this more informative information to improve predictions. Figure 3b shows the attention layer in the neural model architecture. We add the attention layer after the LSTM layer and the attention layer is computed as follows: $\alpha = softmax(W_1 tanh(W_0 M + b_0) + b_1)$, where $M \in \mathbb{R}^m$ is the output from LSTM layer, m is number of nodes in the output layer from LSTM, $W_0, W_1 \in \mathbb{R}^{m \times m}$ are weights, and $b_0, b_1 \in \mathbb{R}^m$ are bias terms. This model has been shown to be effective on social media classification tasks such as mental health prediction [34], rumor detection [25] and gender inference [44].



(a) BERT model



(b) Attention based BERT model

Figure 3: Deep learning model architecture

5 EXPERIMENTAL SETUP

This section describes our experimental setup. We start by presenting details about the data sets used in the two case studies (Section 5.1). We then discuss data issues associated with these two data sets and preprocessing strategies (Section 5.2). Finally, we present our evaluation measures for both performance and fairness of our models (Section 5.3).

5.1 Data set

Our two data sets connect survey responses to Twitter/X accounts, allowing us to have accurate ground truth demographics and accessible Twitter/X data. For both surveys, respondents consented to allowing us to link survey responses to their Twitter/X data to support social science and computer science research. This work is completed under IRB STUDY00003571 and STUDY00002133 at Georgetown University. We will use the first data set to predict gun ownership and the second to predict the willingness of people to receive a Covid-19 vaccine.

Gun ownership. This data set is a nationally representative sample of 2563 individuals. Among all individuals, 661 respondents from the sample consented to having their survey response linked to their Twitter/X data for a two year period. Survey responses included respondent gun ownership status, demographic information, and Twitter/X handles [28]. We use the Twitter/X handles to collect the tweets, biography, and account information of the consented survey respondents. After data collection, we exclude individuals without valid Twitter/X handles, those having limited Twitter/X activity (too much missing data), or individuals who did not provide their gender or gun ownership status. After removing those individuals, we have 541 individuals in the data set for the prediction task. For this case study, the prediction task is to determine whether the user is a gun owner. The sensitive feature is gender. Table 2a shows the distribution of gender and gun ownership in this sample. Within our sample, men are approximately 5% more likely to own guns than women. We will refer to this data set as the Gun Ownership Data Set.

Table 2: Sample distribution for gun ownership and Covid-19 vaccine hesitancy data sets

(a) Gun ownership

Gender	Gun ownership	Count
Female	No	173
Female	Yes	63
Male	No	208
Male	Yes	97

(b) Covid-19

Gender	Willingness to take a vaccine	Count
Female	No	44
Female	Yes	200
Male	No	29
Male	Yes	200

Covid-19. This data set is part of the MOSAIC (Measuring Online Social Attitudes and Information Collaborative) project, a collaboration between SSRS, Georgetown University, and the University of Michigan. It is a nationally representative data set of 9544 individuals, 689 of whom consented to allowing us to link their survey responses to their Twitter/X data to support social science and computer science research. In the survey, we collect respondents opinions on the Covid-19 vaccine, including whether or not they are likely to get vaccinated, their demographic information, and their Twitter/X handles. We use the provided handles to collect account information and tweets from Twitter/X. We then remove individuals who have missing values in their Covid-19 responses or have too few tweets. This leaves 473 individuals in our data set for the prediction task. Using this sample, our goal is to predict whether or not an individual is planning to take (or has already taken) a Covid-19 vaccine. Gender is the sensitive attribute. Table 2b shows the distribution of gender and vaccine willingness. In this data, female respondents are 5.3% more likely not to take a Covid-19 vaccine than male respondents. We will refer to this data set as the Covid-19 data set since we are predicting vaccine hesitancy.

5.2 Data Issue

In Section 3.2, we discussed the prevalence of different types of data issues that arise when using social media data for prediction. Missing values are a prevalent data issue in social media data. Missing values arise in multiple ways: 1) users (subgroups of users) who do not have accounts on specific social media platforms, and 2) users who share limited information e.g., users who have very few posts. In Section 5.1 we mentioned that we removed respondents who post too few tweets. We empirically determined that less than 10 tweets is too few when considering model performance in terms

of accuracy and F1 score. In this section, we want to explore the impact of missing values in terms of Twitter/X activity level on fairness. For example, if female users post fewer tweets then male users, and we remove a large number of female users from our sample, we may introduce bias in the pre-processing.

We define Twitter/X activity as a continuous variable that measures the number of tweets a user posts. Figure 4 shows Twitter/X activity in the gun ownership data set. The x-axis represents the Twitter/X activity and the y-axis shows the number of respondents with at least that level of Twitter/X of activity (sample size), the gender ratio and the gun ownership ratio, respectively. Figure 4a shows that the size of our sample of respondents decreases significantly as the Twitter/X activity increases (from 659 with at least 1 post to 481 with at least 50 post). Figure 4b and 4c show the relationship between tweet activity, the sensitive attribute and the label. Gender ratio is defined as number of male users divided by the number of female users and gun ownership ratio is the number of gun owners divided by number of non gun owners. From the two figures, both the gender and gun ownership ratios remain fairly stable across Twitter/X activity levels, particularly when the activity level is greater than 10. When the Twitter/X activity level is between 10 and 50, the difference between the maximum and minimum gender ratio is less than 0.1 or 7.3% and for the gun ownership ratio, it is less than 0.03 or 6.9%. From this result, we can conclude that setting Twitter/X activity level to 10, i.e., removing all handles posting less than 10 tweets, will reduce activity bias. We will empirically evaluate the impact of lower levels of 'missing' data by assessing the impact on model performance when the number of posts is above 10, but less than 50. We consider this a *moderate* to *low* level of missingness.

Figure 5 shows a similar set of figures for Twitter/X activity on the Covid-19 vaccine data set. Similar to the gun ownership data set, the sample size changes significantly (from 572 with at least 1 post to 379 with at least 50 post). Gender ratio and Covid-19 vaccine acceptance ratio remain fairly constant. Similar to gun violence, setting the minimum Twitter/X activity to 10 posts reduces the amount of bias in our sample. In general, dropping users with too much missing information will reduce the bias to our sample and improve model performance.

5.3 Evaluation

In our fairness evaluation, we use two metrics to quantify fairness, disparate impact ("p%-rule") [24, 63] and CV [10]. The p%-rule is closest to the legal definition of fairness and it is often used in anti-discrimination laws to measure fairness and discrimination [3]. The p%-rule is defined as:

$$\min\left(\frac{P(\hat{Y}=+|S=1)}{P(\hat{Y}=+|S=0)}, \frac{P(\hat{Y}=+|S=0)}{P(\hat{Y}=+|S=1)}\right)$$

The p%-rule measures the probability ratio of getting a positive outcome between the privileged group and the unprivileged group. When using the p%-rule, the higher the value, the fairer the machine learning model. Generally, if the p%-rule is greater than 80%, or 0.8, the model is considered to be non-discriminatory [6]. Another fairness metric, CV, proposed by Calders and Verwer is similar to p%-rule, but uses the difference instead of the ratio to quantify bias. It is defined as:

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

A low value for CV is an indication of a fair classifier. Compared to the p%-rule, CV does a better job of capturing bias when the probabilities are high. For example, if $P(\hat{Y}=+|S=1) = 0.95$ and $P(\hat{Y}=+|S=0) = 0.8$, the p%-rule value is 0.842 (indicating a fair classifier), but the CV value is 0.15, highlighting the difference between the probabilities. When the probabilities are low, p%-rule has values that better align with expected fairness values. For example, if $P(\hat{Y}=+|S=1) = 0.2$ and $P(\hat{Y}=+|S=0) = 0.05$, CV is still 0.15, but p%-rule is 0.25.

In our experiments, we use 5-fold cross validation to train our models and use accuracy and F1 score on the validation set in each fold to measure accuracy and F1, We use p%-rule and CV score to measure fairness. The fairness metrics require that we define the positive and negative outcomes for the sensitive attribute. This means we have to specify the privileged and unprivileged groups. Unlike most classic fairness data sets such as the COMPAS recidivism data and the German credit data [21, 36], there is no clear definition of positive and negative outcome for our data sets.

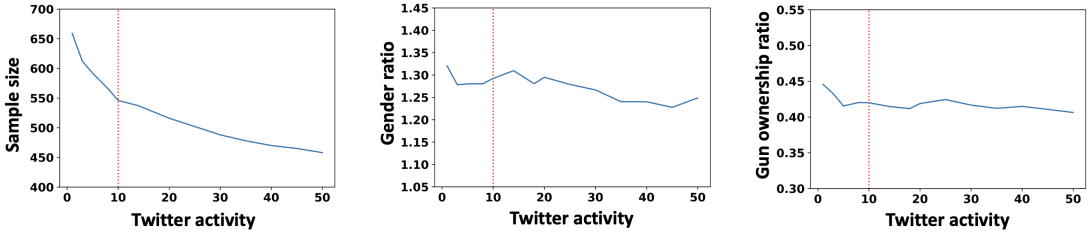
6 EMPIRICAL EVALUATION

In this section, we begin by investigating the fairness of different types of machine learning models for our two cases (Section 6.1). When the classifiers are not fair, we show how fairness can be improved using our fixing approach (Sections 6.2 and 6.3). Finally, we measure and analyze the decrease in relationship between the sensitive feature proxy X^p and the hidden sensitive feature S (Section 6.4).

6.1 Prediction results for case studies

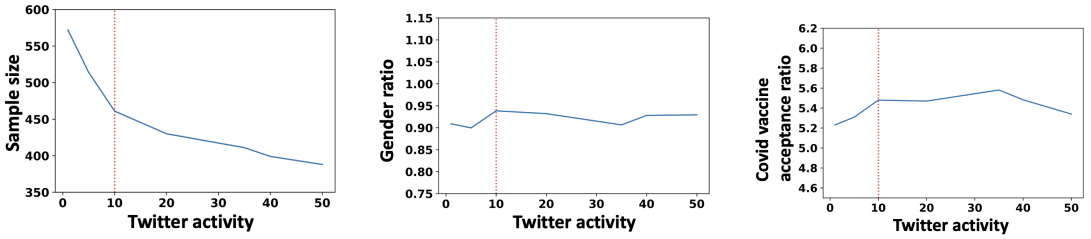
Gun ownership. Table 3 shows the average and standard error (in parentheses) for accuracy, F1 score, and the fairness measures for all the classic machine learning models and neural network models in our study. In terms of model performance, logistic regression has the best overall performance (72%) and the best performance for both the privileged group (77%) and the unprivileged group (67%), followed by the neural network model with an attention layer (69%). Overall, the performance of logistic regression is 3% higher in accuracy and 8% higher in F1 score when compared to the second best model. We believe that the logistic regression model has better performance than the neural network model because our data set is relatively small with approximately 500 observations. Typically, more examples are necessary to train a neural model. In terms of fairness, the random forest and neural network with attention layer models have the best fairness scores and logistic regression has an average fairness score when compared to the other models. The p%-rule score for logistic regression is about 20% lower than random forest, 9% higher for the CV score and the difference in accuracy between the privileged and unprivileged group is 5% less than random forest. However, random forest performs poorly in terms of accuracy (the 2nd lowest) and F1 score (the lowest).

Covid-19. Table 4 shows the model performance and fairness scores for the Covid-19 data set. Similar to the gun ownership data, the neural network models have worse performance than



(a) Number of handles (b) Gender ratio (c) Gun ownership ratio

Figure 4: Twitter/X activity of respondents in the gun ownership data set



(a) Number of handles (b) Gender ratio (c) Vaccine acceptance ratio

Figure 5: Twitter/X activity of respondents in Covid-19 data set

the classic N-gram models. Again, this likely occurs because the sample size is fairly small (462 observations). In terms of model performance, random forest has the best performance, followed by the neural network model with an attention layer and logistic regression. Overall, the performance of random forest is 3% higher than the next best model in terms of accuracy and F1 score. In terms of fairness, the fairness scores are fairly consistent across all models. SVM has the highest fairness score followed by logistic regression. The p%-rule score for random forest is approximately 9% lower than SVM, less than 1% higher than the CV score and the difference in accuracy is 5% higher than SVM. However, SVM has the lowest accuracy and a F1 score when compared to the other models.

For both of our case studies, we find that the best performing models have reasonable accuracy and F1 scores. However, their fairness scores are poor, meaning that they are biased with respect to gender. Therefore, our next step is to determine whether or not we can improve the fairness while maintaining predictive accuracy.

6.2 Fixing using the ground truth sensitive attribute (S)

In this section, we show that if a sensitive attribute S is available, we can use an existing fairness fixing method ([58]) to reduce the correlation between X^p and S, thereby reducing the bias. Figure 6 shows the performance and fairness scores of the best classifier for each data set before and after fixing. The red dots show the performance and fairness scores of the best classifier (using the results from Table 3) for each data set before fixing, the purple dots show prediction results of the best classifier after fixing using the ground truth sensitive attribute. In both data sets, the fairness scores, p%-rule and CV, improve substantially with only a small tradeoff in accuracy and F1 score. For example, in the gun ownership data set, the accuracy decreases by less than 2% and the F1-score by less than 3%, while the p%-score increases to 0.89 (23% increase) and the CV score decreases to 0.05 (8% decrease). Similar to many other studies performed on different types of data sets, these results are a strong indication that when ground truth demographic features are available, researchers can easily use existing fairness fixing methods to reduce bias. But what about when demographic features are not available? We consider this scenario next.

6.3 Fixing method with inferred sensitive attribute ($\hat{S}$)

In this section, we explore inferring the sensitive attribute, i.e. computing $\hat{S}$ for cases when ground truth sensitive feature is not available to the researcher.

Table 5 shows the performance of predicting gender using three state of the art inference models: Siamese, BERT and BERT emoji [43, 44] on the gun ownership and Covid-19 data sets. We note that the three gender inference models we consider in this paper infer binary gender. We acknowledge that this is a limitation since there are more than two possible values for gender. However, in our data sets, we only have binary gender information as ground truth data and therefore, can only infer binary gender from the inference models. Among the three models, BERT emoji has the best performance with an overall accuracy of 0.809 and 0.787 on gun ownership and Covid-19, respectively.

Next, we use the inferred sensitive attributes $\hat{S}$ from each gender inference model as a proxy for the ground truth sensitive attribute S. We then apply the same fixing method to remove selection bias.

These results are also shown in Figure 6. Recall that the left subfigures show the accuracy (x-axis) and F1 scores (y-axis) for the original data (red square), the ground truth data (purple triangle)

Table 3: Results for predicting gun ownership

Model	F1 score	p%-rule	CV	Privileged group accuracy	Unprivileged group accuracy	Overall accuracy
Logistic regression	**0.699 (0.04)**	0.648 (0.061)	0.127 (0.062)	**0.768 (0.033)**	**0.671 (0.028)**	**0.72 (0.03)**
Random forest	0.498 (0.066)	**0.877 (0.067)**	**0.033 (0.029)**	0.662 (0.027)	0.601 (0.026)	0.635 (0.026)
Decision tree	0.481 (0.088)	0.677 (0.105)	0.117 (0.079)	0.642 (0.039)	0.569 (0.033)	0.601 (0.036)
SVM	0.558 (0.041)	0.376 (0.103)	0.216 (0.095)	0.706 (0.019)	0.578 (0.013)	0.646 (0.015)
NN without attention layer	0.588 (0.042)	0.683 (0.098)	0.114 (0.073)	0.684 (0.043)	0.617 (0.04)	0.651 (0.041)
NN with attention layer	0.618 (0.048)	0.773 (0.104)	0.087 (0.064)	0.716 (0.042)	0.653 (0.046)	0.689 (0.044)

Table 4: Results for predicting Covid-19 vaccine hesitancy

Model	F1 score	p%-rule	CV	Privileged group accuracy	Unprivileged group accuracy	Overall accuracy
Logistic regression	0.658 (0.059)	0.744 (0.051)	0.176 (0.043)	0.697 (0.053)	0.599 (0.047)	0.654 (0.05)
Random forest	**0.684 (0.042)**	0.675 (0.064)	0.217 (0.058)	**0.747 (0.051)**	**0.632 (0.062)**	**0.697 (0.041)**
Decision tree	0.574 (0.062)	0.642 (0.068)	0.262 (0.062)	0.658 (0.051)	0.562 (0.055)	0.618 (0.045)
SVM	0.523 (0.074)	**0.761 (0.069)**	**0.158 (0.032)**	0.599 (0.055)	0.532 (0.049)	0.568 (0.053)
NN without attention layer	0.654 (0.03)	0.662 (0.089)	0.227 (0.053)	0.714 (0.035)	0.632 (0.031)	0.671 (0.032)
NN with attention layer	0.676 (0.023)	0.651 (0.098)	0.216 (0.058)	0.731 (0.028)	0.629 (0.024)	0.688 (0.026)

Table 5: Gender inference model performance

(a) Gun ownership

Ground truth	Siamese	BERT	BERT emoji
Male	0.818	0.747	0.827
Female	0.775	0.799	0.785
Overall	0.8	0.769	0.809

(b) Covid-19

Ground truth	Siamese	BERT	BERT emoji
Male	0.79	0.784	0.805
Female	0.766	0.737	0.771
Overall	0.778	0.759	0.787

and each model using the resampling fixing method with the inferred sensitive attribute values (circles) on our two data sets. The right subfigures show the p%-rule (x-axis) and the CV (y-axis) for the same data. From the figure, we see that the accuracy and F1 scores remain high (less than a 2% difference). The fairness scores using $\hat{S}$ are significantly better than without fixing (red square), but not as strong as when using the ground truth sensitive attribute (purple triangle). In the gun ownership data, with inferred sensitive attribute, the p%-rule improves by 18% and the CV score improves by 4.7%. In the Covid-19 data, the p%-rule improves by 10.3% and the CV score improves by almost 6%. A good machine learning model should be in the top right corner of the left subfigures and a fair model should be in the bottom right corner of the right subfigures. From the figures, the red dots, results without any fixing, have good model performance but bad fairness. The purple dots, results with fixing using ground truth gender, have lower accuracy but better fairness, highlighting that fairness can be improved with a small tradeoff with accuracy, even when ground truth data are not available.

6.4 Analyzing the relationship between X^p and S

In Section 3, we suggested that when the sensitive attribute proxy X^p is highly correlated to S, the model may be less fair. The amount of correlation is directly related to the fairness of the machine learning model. Therefore, our goal is to make sure the correlation is low.

To better understand the existing relationship between the sensitive feature proxy X^p and the hidden sensitive feature S, we want to measure the size of the overlap in features between them. To measure the initial overlap, we build a model to predict the sensitive attribute (gender) using the features we used to predict gun ownership. More specifically, we use a set of features to train a logistic regression classifier to predict gun ownership. We then use the same set of features to train another logistic regression model to predict the sensitive attribute. After training both classifiers, we identify the top n most predictive features from both models and determine the overlap, i.e. the number of features in common among the top n most predictive features in the two classifiers. The more features in common, the stronger the relationship between X^p and S, i.e. the more related they are.

Figure 7 shows the proportion of overlapping features in both the gun ownership and the Covid-19 data sets for n between 20 and 60. The proportion is largest when no fixing takes place (red lines) and lowest after fixing using ground truth data (purple lines). When fixing using an inferred gender value, the proportion of overlapping features is in between the the two. We see from Figure 7 that the size of the overlap between the sensitive feature proxy and the hidden sensitive feature reduces by approximately 50% after fixing for all the cases, further supporting the idea that fairness can be achieved with only a small loss in accuracy whether or not the sensitive attribute is available to the researcher.

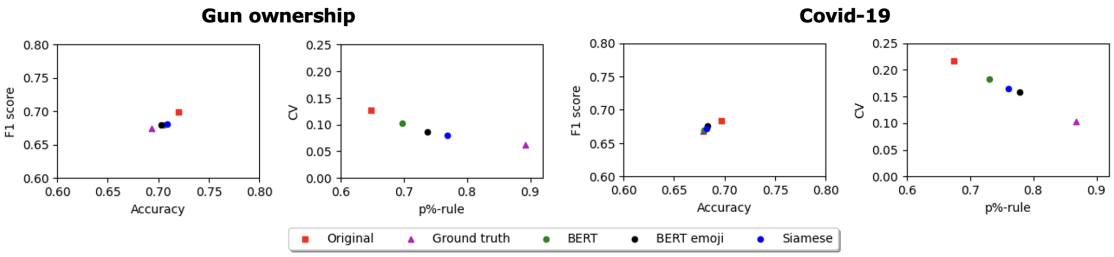


Figure 6: Performance and fairness results before and after fixing with ground truth and inferred sensitive attribute

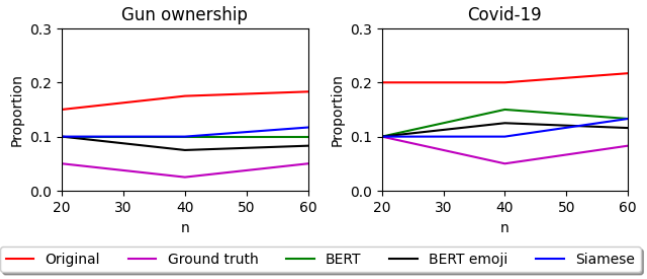


Figure 7: Proportion of overlapping features in gun ownership and Covid-19 data set with and without apply the fix method

7 ETHICS AND REPRODUCIBILITY STATEMENT

In our case studies, we obtained consent from survey respondents to link their survey responses to their Twitter/X accounts. We also received IRB approval (STUDY00003571 and STUDY00002133) from Georgetown University and we use a strict protocol for storing the survey and Twitter/X data from respondents.

We recognize the ethical complexity associated with the use of human trace data for research and have focused on designing an experiment that advances our knowledge of fairness on social media while maintaining the privacy of those who consented to being part of our research. We acknowledge that the detection of user demographics also poses unique ethical considerations. While automated methods can be valuable, error does exist in these models and there are possible equity and justice related consequences to imbalances in these errors. It is one reason that studies like these are important for computer science researchers to conduct and share.

8 CONCLUSION

As computer scientists try to understand opinion and behavior on social media, it is imperative to build fair models in the presence of noisy, missing data. This work presents a methodology for determining the fairness of models built using text social media data when ground truth sensitive attributes are available and in the case when they are not. Using two Twitter/X case studies, we highlight the bias that exists when applying state of the art machine learning models on social media data. This bias is a direct result of the presence of a sensitive feature proxy that correlates to the hidden sensitive feature. We show that when we have the ground truth sensitive feature, we can effectively improve fairness by adapting existing fixing methods. We then show that when we do not have the ground truth demographic feature of interest, we can use existing reliable demographic inference models with high accuracy to infer the sensitive feature and use the inferred feature to improve fairness. This work is a first step toward understanding how to effectively measure fairness in a noisy environment like social media.

There are some important limitations of this work that may suggest future research directions. First, our sample sizes were small. Using self supervised learning or transfer learning may have improved the accuracy of these models and given insight into the fairness of these newer models. We also focused on binary classification tasks. An important future direction is to understand how well these finding hold for multi-class problems. Because we use social media data, we have other types of features available to us, e.g. relationship data and image data. Future work should consider using all these types of data together to see what new fairness issues may arise. Finally, our study was conducted on Twitter/X. It would be informative to conduct similar experiments using data from other social media platforms.

ACKNOWLEDGMENTS

This research was funded by National Science Foundation awards #1934925 and #1934494, the National Collaborative on Gun Violence Research (NCGVR), and the Massive Data Institute (MDI) at Georgetown University. We thank our funders for supporting this work.

REFERENCES

[1] Shakeel Ahmad, Muhammad Zubair Asghar, Fahad M Alotaibi, and Irfanullah Awan. 2019. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences* 9, 1, 1–23.

[2] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Conference on Fairness, Accountability, and Transparency*. 116–128.

[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104, 671.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5, 4–1.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. arXiv:1706.02409

[6] D. Biddle. 2005. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Gower.

[7] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn:*

A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29.

[9] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *conference on machine learning*. 803–811.

[10] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2, 277–292.

[11] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Conference on Neural Information Processing Systems*. 3995–4004.

[12] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* 3, 1, 1–11.

[13] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Conference on fairness, accountability, and transparency*. 339–348.

[14] Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. 2015. A comparative study of demographic attribute inference in twitter. In *Conference on Web and Social Media*, Vol. 9.

[15] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. arXiv:1810.08810

[16] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of twitter users. In *Conference on Social Computing*. 192–199.

[17] Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. 2023. Auditing fairness under unawareness through counterfactual reasoning. *Information Processing & Management* 60, 2, 103224.

[18] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Conference on Fairness, Accountability, and Transparency*. 525–534.

[19] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805

[21] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*. 214–226.

[23] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability and Transparency*. 172–186.

[24] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Conference on Knowledge Discovery and Data Mining*. 259–268.

[25] Yue Geng, Zheng Lin, Peng Fu, and Weiping Wang. 2019. Rumor detection on social media: A multi-view model using self-attention mechanism. In *Conference on Computational Science*. 339–352.

[26] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Conference on Knowledge Discovery and Data Mining*. 2221–2231.

[27] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When fair ranking meets uncertain inference. In *Conference on Research and Development in Information Retrieval*. 1033–1043.

[28] Carole Roan Gresenz, Lisa Singh, Yanchen Wang, Jaren Haber, and Yaguang Liu. 2023. Development and Assessment of a Social Media–Based Construct of Firearm Ownership: Computational Derivation and Benchmark Comparison. *Journal of medical internet research* 25, e45187.

[29] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, Vol. 1. 2.

[30] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18, 43–49.

[31] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29, 3315–3323.

[32] Hui Hu, Mike Borowczak, and Zhengzhang Chen. 2021. Privacy-Preserving Fair Machine Learning Without Collecting Sensitive Demographic Data. In *Conference on Neural Networks*. 1–9.

[33] Adam Hughes and Stefan Wojcik. 2019. Key takeaways from our new study of how Americans use Twitter. https://www.pewresearch.org/fact-tank/2019/04/24/key-takeaways-from-our-new-study-of-how-americans-use-twitter

[34] Julia Ive, George Gkotsis, Rina Dutta, Robert Stewart, and Sumithra Velupillai. 2018. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. 69–77.

[35] Kokil Jaidka, Saifuddin Ahmed, Marko Skoric, and Martin Hilbert. 2019. Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication* 29, 3, 252–273.

[36] Lauren Kirchner Jeff Larson, Surya Mattu and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm.

[37] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1, 1–33.

[38] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *European Conference on Machine Learning and Knowledge Discovery in Databases*. 35–50.

[39] Akrivi Krouska, Christos Troussas, and Maria Virvou. 2020. Deep Learning for Twitter Sentiment Analysis: The Effect of Pre-trained Word Embedding. *Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications*, 111–124.

[40] Susan Leavy. 2018. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *International Workshop on Gender Equality in Software Engineering*. 14–16.

[41] Nicol Turner Lee. 2018. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*.

[42] Sabina Leonelli, Rebecca Lovell, Benedict W Wheeler, Lora Fleming, and Hywel Williams. 2021. From FAIR data to fair data use: Methodological data fairness in health-related social media research. *Big Data & Society* 8, 1.

[43] Yaguang Liu and Lisa Singh. 2021. Age Inference Using A Hierarchical Attention Neural Network. In *Conference on Information & Knowledge Management*. 3273–3277.

[44] Yaguang Liu, Lisa Singh, and Zeina Mneimneh. 2021. A Comparative Analysis of Classic and Deep Learning Models for Inferring Gender and Age of Twitter Users. In *Conference on Deep Learning Theory and Applications-DeLTA*.

[45] Ninareh Mehrabi, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Debiasing community detection: The importance of lowly connected nodes. In *Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 509–512.

[46] Nor Rahayu Ngatirin, Zurinahni Zainol, and Tan Lee Chee Yoong. 2016. A comparative study of different classifiers for automatic personality prediction. In *Conference on Control System, Computing and Engineering*. 435–440.

[47] Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. 2021. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Uncertainty in Artificial Intelligence*. 600–609.

[48] R Colby Perkins. 1993. *Evaluating the Passel-Word Spanish surname list: 1990 decennial census post enumeration survey results*. US Department of Commerce, Economics and Statistics Administration.

[49] Andrew Perrin. 2015. Social media usage. *Pew Research Center* 125, 52–68.

[50] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. 2021. Post-processing for Individual Fairness. *Advances in Neural Information Processing Systems* 34.

[51] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30.

[52] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084

[53] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. arXiv:1904.05233

[54] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin EP Seligman. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8, 9.

[55] Eli Sherman, Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models. In *Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. 217–223.

[56] Lisa Singh, A Polyzhou, Yanchen Wang, Jason Farr, and C Gresenz. 2020. Social Media Data-Our Ethical Conundrum. *Bulletin of the IEEE Computer Society*

Technical Committee on Database Engineering 43, 4.

[57] Vladimir Vapnik, Rauf Izmailov, et al. 2015. Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research* 16, 1, 2023–2049.

[58] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 1–19.

[59] Yifang Wei and Lisa Singh. 2017. Using network flows to identify users sharing extremist content on social media. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 330–342.

[60] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. 2019. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 1–1.

[61] Marilyn A Winkleby and Beverly Rockhill. 1992. Comparability of self-reported Hispanic ethnicity and Spanish surname coding. *Hispanic Journal of Behavioral Sciences* 14, 4, 487–495.

[62] Runhua Xu, Nathalie Baracaldo, and James Joshi. 2021. Privacy-preserving machine learning: Methods, challenges and directions. arXiv:2108.04417

[63] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.

[64] Yan Zhang. 2018. Assessing fair lending risks using race/ethnicity proxies. *Management Science* 64, 1, 178–197.