Data Mining and Knowledge Discovery (2022) 36:2194–2213
https://doi.org/10.1007/s10618-022-00855-y

Fairness in vulnerable attribute prediction on social media

Mariano G. Beiró[1,2] · **Kyriaki Kalimeri**[3]

Received: 30 August 2021 / Accepted: 5 July 2022 / Published online: 17 September 2022

Abstract
Historically, policymakers and practitioners relied exclusively on survey and census data to design and plan for assistive interventions; now, social media offer a timely and cost-effective way to reach out to populations otherwise unobserved. This study was designed to address the needs of a non-for-profit organisation to reach out to the young unemployed individuals in Italy with educational and job opportunities via communication channels that are more likely to appeal to younger generations. To this extend, we developed an ad-hoc Facebook application which administers questionnaires while gathering data about the Likes on Facebook Pages. Then, we developed a machine learning framework that successfully predicts the unemployment status of an unseen individual (.74 AUC). However, blindly delegating to the machine learning model the communication intervention may lead to digital discrimination on the basis of socio-demographic characteristics. Here, we propose a framework that aims to optimising both for the prediction performance as well as the most adequate fairness metric. Our framework is based on an adaptive threshold for gender, while we show that it can be expanded for other socio-demographic attributes and generalised for other interventions of assistive character. We present a doubly cross-validated setting that achieves out-of-sample stability and generalisability of results. We compare the behaviour of models that infer on different sets of data and provide an indepth discussion on the most predictive features, demonstrating that the "fairness through unawareness" approach does not suffice to achieve a fair classification since sensi-

Responsible editor: Toon Calders

✉ Mariano G. Beiró
mbeiro@fi.uba.ar

✉ Kyriaki Kalimeri
kyriaki.kalimeri@isi.it

1 Facultad de Ingeniería, Universidad de Buenos Aires, Av. Paseo Colón 850, Buenos Aires, Argentina

2 CONICET–Universidad de Buenos Aires, INTECIN, Av. Paseo Colón 850, Buenos Aires, Argentina

3 ISI Foundation, Turin, Italy

tive demographic information can be inferred not only via other sociodemographic attributes but also from behavioural digital patterns. Finally, we thoroughly assess the behaviour of the adaptive threshold approach and provide an in-depth discussion on the advantages but also the implications of such models offering actionable insights. Our results show that careful assessment of fairness metrics should be considered, primarily when AI models are employed for policymaking.

Keywords Social Media · Unemployment · Fairness · Machine learning · Digital discrimination · Data for social good

1 Introduction

Youth unemployment is one of the most significant challenges modern societies are facing. Direct consequences include poverty, social exclusion, and criminal behaviours, while negative impact on the future employability and wage cannot be obscured. In EU more than 3 million young people are unemployed, while in 2014, youth unemployment in Italy reached 46% (ISTAT 2020). Today, this rate is at 37% (ISTAT 2020), substantially lower than 2014, but still in alerting levels and well above the average European trend[1].

Policymakers and practitioners strive to reach out to the vulnerable populations of interest in the most efficient way possible. This study was designed exactly to address the needs of an Italian nonprofit organisation to reach out to the young unemployed individuals with educational and job opportunities, via communication channels that are more likely to appeal to younger generations. To this extent, we developed an ad-hoc Facebook-hosted app whose main functionality was to administer a series of questionnaires while at the same time gauging the participants' Likes on Facebook Pages. Combining survey data with social media data, and in particular likes on Facebook Pages, we framed our design as an unemployment classification task. We trained a machine learning (ML) classifier that, based solely on the Likes on Facebook Pages, successfully predicts the occupational status of new individuals that did not provide any demographic information. Those classified as "unemployed" would then be presented with the respective communication.

Social media are now a valid, complementary tool to get actionable insights on populations that are hard to reach or even timely deploy a communication plan (Kalimeri et al. 2020). The plentifulness of social media data combined with artificial intelligence (AI) methods for processing have a huge impact in interdisciplinary fields such as social sciences and humanities which continues to grow exponentially (Ntoutsi et al. 2020). However, as all sources of data come with their own biases and limitations (Olteanu et al. 2016), hence caution is needed when applied to social sciences, to avoid generating or even amplifying existing social inequalities. Traditionally, in computer science the performance score would be the only criterion according to which a machine learning (ML) model would be optimised. In several domains including social sciences though, decisions driven by AI-based insights may treat individuals

[1] https://ec.europa.eu/social/main.jsp?catId=1036

differently, based on their personal characteristics (O'Neil 2016) such as gender, age, or other attributes, or even unfairly.

The concept of "fairness" is inherently subjective and can have different interpretations and definitions depending on the specific problem under investigation (Verma and Rubin 2018). Here, since our interventions are meant to have an assisting character, our definition of fairness focuses on *parity of opportunity*. This means that our focus is on avoiding disproportionally missing individuals from certain sociodemographic groups such as some gender, age group, or geographical region, among those who would potentially be entitled to receive the communication, hence the benefit. We focus on automatically identifying the unemployed population as inferred from social media traces, placing the focal point on accurate yet "fair" predictions. Hence, our mission goes well beyond the creation of an accurate machine learning model. The questions we ask ourselves are: *do our models introduce any discrimination?* And if yes, *can we account for it?* and *how harmful would this be?* Overall, *are our predictions "fair" enough?*

To answer these questions, we build a series of ML models, assessing the predictive power of digital and demographic data both in terms of accuracy (AUC[2]) and fairness. Although, our ML models achieve a state-of-the-art performance in terms of accuracy automatic prediction of the employment status, we dive further into highlighting the biases and trade-offs when introducing the notion of fairness expressed as the equality of opportunity. We also discuss the limitations of the simple "fairness through unawareness" approach, according to which fairness can be achieved when the model simply ignores all protected attributes, i.e. gender, age, and region.

Here, we build on the seminal work of Hardt et al. (2016), introducing an adaptive threshold criterion on a real-life case study scenario. We thoroughly discuss the potentials and limitations of this approach, showcasing its generalisability in other demographic data as well as different configurations. Importantly, in line with evidence from the current literature (Kalimeri et al. 2019; Pedreshi et al. 2008), we confirm that exclusion of protected attributes from the ML model's set of predictors does not, per se, guarantee a "fair" model since socio-demographic attributes might be embedded in our behavioural digital patterns. Finally, we provide key observations aiming to help practitioners generalise this approach to other domains, for instance for humanitarian crisis management, where AI models are increasingly more involved in the decision-making process (Aiken et al. 2022).

2 Related work

Digital data from web queries have been employed to predict unemployment rates since more than a decade (Gao et al. 2019). More recently, other digital traces were employed to infer the unemployment trends. Toole et al. showed that mobile phone activity patterns revealed valuable indicators of the socio-economical status of geographical regions (Toole et al. 2015). At the same time, changes in the calling behaviours were also found useful when forecasting macro unemployment rates (Sundsøy et al. 2016).

[2] We conventionally refer to the AUROC values as "accuracy" throughout this paper.

Social media data were also employed in the fight against unemployment. The Twitter platform proved particularly useful in nowcasting and forecasting unemployment rates (Bokányi et al. 2017; Llorente et al. 2015); diversity in mobility fluxes, diurnal rhythms, and grammatical styles were associated with employment status.

Given its broad population penetration worldwide, together with the possibility to administer targeted communications, the scientific community is increasingly more employing data from the Facebook platform to assess social phenomena (Kalimeri et al. 2020). Facebook advertising data have been employed to monitor poverty (Fatehkia et al. 2020) and social inequalities (Fatehkia et al. 2018; Rama et al. 2020). The advertising platform is also shown to have great potentials in reaching vulnerable and otherwise hard-to-reach segments of the population, since it allows to target individuals based not only on their basic demographic data but on behaviours and preferences (Eslami et al. 2018). More recently, (Urbinati et al. 2020) employed Facebook data together with personality and moral values to further explore psychological and cultural differences between the employed and unemployed communities in Italy.

The labour market has long suffered from discrimination (Becker 2010), with literature providing evidence of unfair treatment in personnel selection (Stoll et al. 2004) and wages (Kuhn 1987) as a result of race and sex discrimination, respectively. As increasingly more decisions are delegated to algorithms, digital discrimination is becoming an important issue (Yeung and Lodge 2019; Barocas and Selbst 2016). This discrimination can arise either from bias in the data or in the algorithms. In the former, it might be the case that the sample is not representative of the target population, or that it reflects historical discrimination patterns which the algorithm will replicate. In the latter, the objective function used by the algorithm might prioritize the correct classification of users in the majority class, producing unfair results. To measure the implications of these biases, different notions of fairness have been proposed, such as *demographic parity* (Calders and Verwer 2010), *calibration* (Kleinberg et al. 2016), *equality of opportunity* or *equalized odds* (Hardt et al. 2016). However, it has been shown that it is not possible in general to satisfy many of these notions at the same time (Corbett-Davies et al. 2017; Dutta et al. 2020; Kleinberg et al. 2016). In these cases, the correct measure to choose will depend on the application requirements. In the context of clustering tasks, specific notions of fairness have been proposed, as covered in Chhabra et al. (2021). Due to the social implications of unfairness many claim that people should be involved in the datasets' construction process through participatory mechanisms, to ensure data inclusivity (Akintande 2021), and that good research practices and methodologies should be developed to foster distributive fairness (Leonelli et al. 2021). Logistically, such approach is not always feasible.

In machine learning, increasingly more studies address the topic of algorithmic fairness from various perspectives which can be grouped in three major categories; pre-processing, in-processing, and post-processing (see Pessach and Shmueli 2022 for a comprehensive review focusing on classification tasks). Pre-processing approaches include learning a representation that obfuscates the sensitive information (Zemel et al. 2013), or suppressing the sensitive attribute, changing some labels, or sampling (Kamiran and Calders 2012). In-processing techniques encompass using minimax optimization (Agarwal et al. 2018), or introducing constraints (Zafar et al. 2017) and regularization terms (Kamishima et al. 2012) into the objective function to

set the trade-off between fairness and accuracy. Finally, post-processing techniques include flipping some decisions or using different thresholds to achieve fairness (Hardt et al. 2016). When some assumptions can be made about the data generating process, causal reasoning as proposed in Kilbertus et al. (2017) can help mitigate discrimination of machine learning models.

In the context of labour market, there are still limited applied research case studies regarding fairness aware predictive models (Desiere et al. 2018). In particular, a study from 2020 in Flanders (Belgium) assessed the trade-off between accuracy and equity in AI models for profiling job-seekers at high- or low-risk (Desiere and Struyven 2020) using the labour market trajectory of 288, 000 job-seekers as recorded by the public employment service (VDAB). They showed that statistical discrimination is an inherent feature of AI-based profiling models. Moreover, and as pointed by the same authors (van Landeghem et al. 2021), these models might even reinforce existent discrimination patterns of unemployment, if not controlled for bias. In a study very close to ours, (Bonanomi et al. 2017), predicted the occupational status and in particular the NEET status of individuals and having information only about their likes on Facebook Pages. However, they did not assess whether the predictions of their models were biased towards a certain class which is one of the points we aim to improve in this study.

Exploring the trade-off between accuracy and fairness, (Dutta et al. 2020), showed that real world data give noisier (and hence biased) mappings for the unprivileged group due to historic differences in opportunity, representation, etc. making their positive and negative labels "less separable". Working on synthetic data, they concluded that it is problematic to measure accuracy with respect to data that reflects bias, and instead, we should be considering accuracy with respect to ideal, unbiased data. Here, we study the trade-off between accuracy and fairness in real-world data, pointing out biases and pitfalls. Finally, inspired by the work of Hardt et al. (2016) who studied the trade-off between false negative and false positive rates, we propose an adaptive threshold approach to address those shortcomings in a real application scenario.

3 Experimental design and data collection

This study was designed to address a request from an Italian non-for-profit organisation aiming to reach out to the unemployed community with educational and job opportunities via the Facebook Platform. Together with the practitioners, we created an ad-hoc Facebook-hosted application whose major function was to administer questionnaires. After entering the app, the participants were asked to provide basic demographic information, namely gender, employment status and the province of residence; however, they were free to proceed without filling in this information. The app was disseminated initially via email invitations sent to the existing cohort of our partners and successively via two nationally-wide traditional media campaign. Upon providing their informed consent, participants agreed to provide us with their "Likes" on Facebook Pages. The application was mainly deployed in Italy and was initially launched in March 2016. The data used here were downloaded in September 2019. The final aim of the practitioners was to discover the unemployed users of the app

Table 1 Demographic breakdown of our data according to gender, age and occupation

	Census	Dataset *n = 11,393*
Gender		
Female	51.1%	38.1%
Male	48.4%	61.8%
Age		
17–24	7.9%	43.1%
25–34	11.0%	31.2%
35–44	13.8%	13.6%
45–54	16.1%	7.1%
55–64	13.3%	4.5%
65+	24.5%	0.3%
Occupation		
Employed	77%	43.9%
Unemployed	8.7%	7.4%
Student	14.2%	48.5%

The "Census" column reports the national distribution per attribute according to the statistics provided by the official census bureau (ISTAT 2020). The "Dataset" column reports the percentages of the total number of participants for which we have complete demographic records

- regardless of them explicitly stating their occupation - and communicate with the educational and job opportunities.

Demographic Information. Out of the 63,980 users that entered our application, only 11,393 provided us with full demographic records (gender[3], age, and occupation) while at the same time had a sufficient number of likes on Facebook Pages (> 50). Table 1 presents a comparison of the official Italian Census for 2019 (ISTAT 2020) and the respective population percentage breakdown in our sample. Our cohort presents a slight over-representation of males and people in the 16-24 age group. In terms of occupation, we notice that we have an over-representation of students. The Lombardy region is over-represented, a phenomenon explained by the fact that the project was initially launched in that region. At the same time, we notice that the area of Marche is under-represented, while all other regions in our cohort follow the distribution of the official Italian Census closely. [4] Gender, age, and geographic biases are expected since the seed population for our study was a youth cohort based in the Lombardy region. For the purpose of our study, slight deviations from a perfectly representative sample are not influencing the results since our definition of fairness relies on addressing equally the individuals of every socio-demographic group and not with respect to their proportion in the population according to the official census. We consider participants who declared their occupation status to be 'student' and 'employed' as one single

[3] The gender attribute is considered to be a binary variable since very few participants opted for the "Other" option.

[4] A comparison between the geographical distribution of our sample per region and the expected values from the official Census is shown in the Supplementary Materials.

Table 2 Description of the original and engineered features in the dataset

Description	# Features
Liked Pages	2,063,944
Liked Pages per Category To express how much a participant is interested in different categories of pages, we compute the number of pages he gave like to inside each Category.	1.553
Normalised Categories: As the participants' activity can greatly vary, we normalise the *Liked Pages per Category* to have sum 1.	1.553
Median Page Popularity This index shows how much a participant likes popular pages. The *popularity* of a Page is the number of users that gave like to it, as reported by Facebook in the Page profile.	1
Standard Deviation of Page Popularity	1
Median Category Popularity This index shows how much a participant likes popular categories.	1
Total number of Page likes One feature containing the total number of pages liked by the participant.	1
Total number of liked Categories One feature containing the total number of categories with pages liked by the participant.	1

group[5]. The unemployed, our population of interest, consist the 7.4% of the total population. When addressing vulnerable or minority populations, it is common to deal with heavily unbalanced classes.

Facebook Data. We collected information about the Facebook Pages "liked" by each participant, together with some necessary metadata about the page, such as the name, category[6], and popularity as provided by the Facebook API indicating how many people have "liked" the specific Page. Worth mentioning is the fact that we do no collect information about the posts or the comments of the page. The participants in our dataset have liked approximately 2 million unique *Pages*; for each one of these, we engineered a series of basic participant activity features based on the aforementioned information (see Table 2).

4 Methods

Learning model. We postulate the study as a supervised classification task. We aim at inferring the *occupational status* (unemployed vs employed) of the participants solely from their Facebook Likes on pages, demographic features (age and gender) and the emerging higher-level features (see Table 2). To assess the effects of "hiding" protected attributes from the learning process, along with the previous model (**Demo**) we also provide a model that excludes these features (**NoDemo**). The assumption

[5] This choice is based on the fact that both groups do not actively search for a job.

[6] Link to the list of categories: https://developers.facebook.com/docs/commerce-platform/catalog/categories/google-product-category-to-facebook-product-category

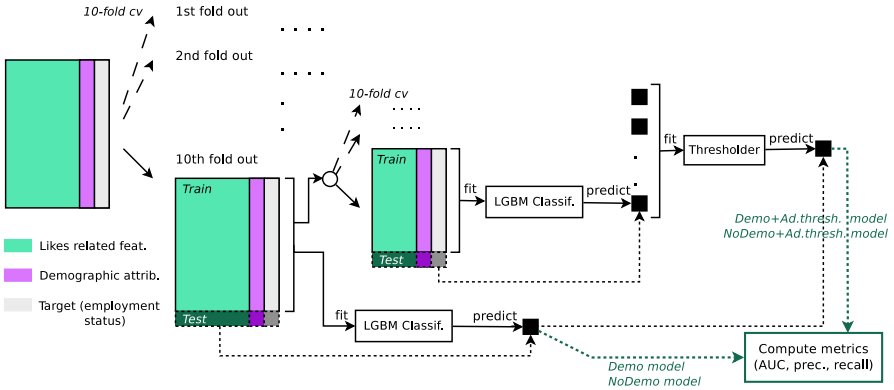

Fig. 1 Cross-validation pipeline for the evaluation of the classification models and thresholds. For the two models with default threshold (Demo and NoDemo) we perform 10-fold cross-validation to compute out-of-sample scores (AUC, precision, recall, and fairness). This is depicted through the LGBM classifier at the bottom. To correctly assess the out-of-sample performance of the models with gender adapted threshold (Demo+AT and NoDemo+AT), we avoid using the same out-of-sample predictions as input to determine the thresholds. Instead, we define inner 10-fold cross-validation inside each fold, whose predictions are used to fit the thresholds. Finally, we test these thresholds by making predictions on the fold that had been left out in their external cross-validation step (Color figure online)

behind this model is that if the algorithm does not explicitly know the demographics of the individual, it will not discriminate against those attributes. Figure 1 depicts the methodological pipeline for the training and testing procedure.

The core of our framework is based on gradient boosting (Ke et al. 2017), and in particular, the *LightGBM* implementation, due to its speed and ability to deal with unbalanced classification scenarios with a large number of features as ours. To avoid overfitting and ensure generalisation, a 10-fold stratified cross-validation scheme was employed for training and validation. Nested cross-validation inside each training set was employed for hyper-parameter optimisation, and the best performing LightGBM model was used for training. The evaluation was performed on the remaining fold so that the classifier was evaluated on data that had never seen before, even during parameter optimisation. The explored hyperparameters were: number of estimators, maximum depth, regularisation, and learning rate [7].

As the dataset is heavily unbalanced, we configured the LightGBM estimator to use balanced class weighting: in this way, the loss function penalises errors with a weight that is inversely proportional to the class sizes. The model performance was assessed in terms of the AUC statistic (Area Under the Receiver Operating Characteristic curve) (Mason and Graham 2002). The AUC was preferred over the commonly-used *accuracy* metric (i.e., the proportion of true positives and true negatives among the total number of samples) as it takes into account the effect of unbalanced labels, which holds true for the occupational status target [8].

[7] The full ranges for each hyperparameter are reported in the Supplementary Materials.

[8] All experiments are performed in Python (Van Rossum and Drake 2009) with scikit-learn (Pedregosa et al. 2011).

Explainability. We employed SHAP (SHapley Additive exPlanations), a game theory approach developed to explain the contribution of each feature to the final output of any machine learning model (Lundberg and Lee 2017a). SHAP values provide both global and local interpretability, meaning that we can assess both how much each predictor and each observation, respectively, contribute to the performance of the classifier. The local explanations are based on assigning a numerical measure of credit to each input feature. Then, global model insights can be obtained by combining many local explanations from the samples (Lundberg et al. 2019). As mentioned by the authors, the classic Shapley values can be considered "optimal" in the sense that within a large class of approaches, they are the only way to measure feature importance while maintaining several natural properties from cooperative game theory (Lundberg and Lee 2017b). SHAP's output helps to understand the general behaviour of our model by assessing the impact of each input feature in the final decision, thus enhancing the usefulness of our framework.

Fairness. As discussed in the introduction, we are interested in assessing and accounting for discrimination concerning specific protected attributes. In the context of Fairness, bias is defined as a disparity measure of a group metric value when compared to a reference group, as described in Saleiro et al. (2018). To ensure that we will not disproportionately miss individuals from specific groups, we focused on Type II parity, i.e., parity of False Negative Rates (FNR). Thus, we define the FNR disparity of a group g from a protected attribute G as:

$$\begin{aligned} \text{FNR}_g \text{ disp.} &= \frac{\text{FNR}_g}{\text{FNR}_{ref.group}} \\ &= \frac{Pr\left[\hat{Y}=0|Y=1 \wedge G=g\right]}{Pr\left[\hat{Y}=0|Y=1 \wedge G=ref.group\right]} \quad , \end{aligned} \tag{1}$$

where Y and $\hat{Y}$ represent the real and predicted target values respectively, and 1 represents the 'unemployed' condition. We formulate disparity using the "80% rule", or $\tau = 0.80$ that controls the range of disparity values that can be considered fair. This notion of parity requires all biases FNR_g disp. to be within the range defined by $(\tau, \frac{1}{\tau})$. We set our reference groups to be the most popular ones for each demographic feature, i.e.: 'Male' for gender, '17-24 years' for age, and 'Lombardy' for geographic region.

To ensure equality of opportunity concerning gender, we extend our original models to include an *adaptive threshold* (Hardt et al. 2016). The default threshold that binary ensemble classifiers have as reference is 0.5 (i.e., predictions with a probability above 0.5 are taken as 'unemployed' while predictions below 0.5 are accepted as 'employed'). Here, we enforce an adaptive threshold to manage the interplay between precision and recall, aiming for equality of opportunity for both genders. We apply the adaptive threshold approach on our two predictive models **Demo** and **NoDemo** obtaining two new models, the **Demo+AT.** and **NoDemo+AT.** models, respectively. Note that the adaptive threshold is not applied directly in the training phase of each model to avoid overfitting. Instead, and as shown in Fig. 1, we fit the threshold including a new nested cross-validation step inside the training. Then we evaluate its performance

out-of-sample. In this way, the fitted threshold will perform equally well on out-of-sample data. Figure 1 provides a visual representation of how the threshold stage is incorporated in our general learning framework.

In the Supplementary Materials we reproduce this adaptive-threshold design at the age level (i.e., obtaining different thresholds for each age-group), and we analyze the performance in that case. Our method can also be employed to guarantee fairness for a combination of demographic features at the same time (for example, gender+age), but we were not able to report results for this setting due to lack of enough samples for some of the (gender, age) groups to run the nested cross-validation procedure.

5 Results & discussion

Learning. Table 3 reports the performances for all the proposed approaches in terms of the AUC, precision, recall, and fairness metrics. As expected when including the age and gender features, **Demo**, the accuracy reaches .74 AUC, outperforming the **NoDemo** model which is agnostic of the demographic information. While both models achieve a clear improvement over the baseline[9] and the current state of the art, when looking at precision and recall we notice that while the recall is quite high, the precision is remarkably low. Hence, the model has a high probability of retrieving relevant samples (unemployed), but the probability of a randomly selected retrieved sample being relevant is low. Practically, for our hypothetical scenario, this means that both models would be very likely to reach many employed individuals too, resulting in a cost-inefficient communication campaign.

Diving deeper into the possible demographic biases of our models, we estimated the AUC score per gender and age group. The aim here is to assess the extent to which our models favour a specific gender or age group while predicting the employment status. Table 3 in the "Demographic accuracy" section, shows that both models, when inferring the employment status, are privileging the females and the 35-44 age group. The practical implications of this are that our hypothetical communication campaign would more likely reach people from those demographic groups. Importantly, the same behaviour emerges for both models, regardless of the inclusion or not of the demographic attributes in the models' learning phase. Such finding shows proof of the claim that just by "hiding" a demographic attribute from the learning process does not result in a bias-free model since behavioural signals may incorporate latent information about these attributes. In fact, in the scientific literature, the gender attribute is shown to be accurately predicted from a series of plain digital behaviours such as web-browsing data, smartphone app usage, search queries, or even simple social networking features (see Table 4).

Explainability. A natural step forward is to assess the most predictive features of each model to evaluate whether the demographic differences are substantially expressed there. We employed the SHAP (SHapley Additive exPlanations) technique to assess the global and local interpretability of the predictors.

[9] The baseline AUC for our tasks is .50.

Table 3 Cross-validated (10-fold) accuracy and fairness results for the prediction of the *Occupational Status* from Likes on Facebook Pages

	Demo	NoDemo	Demo+AT.	NoDemo+AT.
Global Accuracy (Metric: AUC(std))				
Baseline	.50	.50	.50	.50
State of the Art	–	.61(.01) (*)	–	–
Our Approach	.74(.02)	.71(.02)	.74(.02)	.71(.02)
Precision and Recall				
Precision	.16(.02)	.18(.01)	**.26(.05)**	**.25(.03)**
Recall	**.56(.05)**	.48(.02)	.21(.05)	.22(.04)
Demographic accuracy (Metric: AUC(std))				
Gender (M)	.66(.05)	.64(.04)	.66(.05)	.64(.04)
Gender (F)	.78(.02)	.76(.02)	.78(.02)	.76(.02)
Age (17-24)	.70(.08)	.69(.08)	.70(.08)	.69(.08)
Age (25-34)	.66(.05)	.65(.05)	.66(.05)	.65(.05)
Age (35-44)	.74(.09)	.73(.08)	.74(.09)	.73(.08)
Age (45-54)	.61(.17)	.54(.16)	.61(.17)	.54(.16)
Age (55+)	.46(.31)	.46(.29)	.46(.31)	.46(.29)
Fairness (Metric: $\frac{FNR}{FNR_{ref}}$)				
Gender (ref.class: Male)				
Female	.47(.11)	.58(.14)	**1.0(.07)**	**1.02(.14)**
Age (ref.class: 17–24)				
25-34	.35(.08)	.62(.12)	.75(.09)	**.80(.08)**
35-44	.26(.12)	.49(.2)	.71(.09)	.73(.1)
45-54	.41(.24)	**.82(.35)**	**.82(.17)**	**.84(.19)**
55+	.59(.36)	**.99(.42)**	**.82(.18)**	**.91(.19)**

Bold values for precision and recall highlight the best performing models (up to one standard deviation), while bold values for fairness highlight fair models (i.e., those within the range 0.80–1.25)
Demographic accuracy presents the AUC scores for each demographic attribute while predicting the occupation. Bold numbers for fairness point out that average disparity measures were kept in the range $(\tau, \frac{1}{\tau})$. The metric of the reference classes equals 1 and is hence not reported in the Table. We report the distribution for each of the demographic attributes. We consider as "fair" scores that are within the range of .80-1.25 of the Fairness Metric
(*) (Bonanomi et al. 2017)

Figure 2a depicts the contribution of each predictor to the actual output. Each point in this summary plot is a Shapley value for a feature and an instance. The rows in the plot correspond to the most predictive features, while the x-axis value of a point denotes the Shapley value of a specific instance. The Shapley value represents the contribution of the feature value to the unemployed prediction (i.e., positive Shapley values represent that, for this instance, the feature value made the model's prediction tend towards the 'unemployed' label, while a negative value made the model tend towards the 'employed' label). Finally, the colour-coding represents the value of the feature, from low to high. Overlapping points are jittered in the y-axis direction, so we get a sense

Table 4 Related literature on gender(male/female) prediction from digital data

Related Study	Gender (AUC)	Data Source
Present Study	**94%**	**Facebook Likes**
Kosinski et al. (2013)	93%	Facebook Likes
Kalimeri et al. (2019)	90%	Web-browsing & Application Usage
Malmi and Weber (2016)	90%	User Applications
Goel et al. (2012)	85%	Client web-browsing history
Zhong et al. (2015)	85%	Location check-ins
Ying et al. (2012)	85%	Smartphone Call Logs
Bi et al. (2013)	80%	Search Queries
Dong et al. (2014)	80%	Social Networks
Felbo et al. (2017)	79%	Smartphone Call Logs
Seneviratne et al. (2015)	74%	Apps - Category and Content

The bold line represents the AUC for gender prediction obtained in this study
All the above studies reported their findings in terms of AUC metric and are, hence, directly comparable

of the distribution of Shapley values per feature. We observe that for some features, higher values are indicating an 'unemployed' label (e.g., Health/Beauty (cat.norm.) feature), while for others, higher values contribute to the 'employed' label (e.g. Sei allo Scientifico (page)). The features are ordered according to their importance, which is computed as their average impact on the prediction throughout all the samples.

Basic demographic attributes, like age and gender, rank high in the list of important predictors according to the SHAP values, indicating that the model tends to heavily rely on them in its decision-making process. Several of the remaining predominant non-demographic features rank high both for the **NoDemo** and model **Demo** (see Fig. 2b). Overall, this is a desired outcome, however, looking closer at some features we notice alerting behaviours. For instance, the *Health and Beauty* page is a feature that ranks high among the top predictors for both **Demo** and **NoDemo** models. According to Urbinati et al. (2020), a possible interpretation of this is that the unemployed population often follows such pages to get informed about promotions. At the same time, the *Health and Beauty* page is one of the top predictive features of gender (as shown in section E of the Supplementary Materials, by the SHAP values of a gender-prediction model). Such observations contribute to the literature against the "fairness through unawareness" concept, due to the existence of redundant encodings, paving the way to indirectly predict protected attributes from other features (Hardt et al. 2016; Pedreshi et al. 2008). Gender is indeed reflected in many digital behaviours as shown by the related literature (see Table 4), making the task of "gender-fair classification" difficult. Such findings call for attention from scientists and practitioners who should be aware not only about the known data quality and platform issues (Kalimeri et al. 2020; Olteanu et al. 2019) but also about the behavioural patterns tight-knit with sensitive personal attributes. Dutta et al. (2020), in a theoretical approach, demonstrated that this limitation can be overcome by obtaining additional, high quality, data (e.g., status updates Matz et al. 2019). They showed that in this way separability between groups is increased, alleviating the inherent accuracy-fairness trade-off.

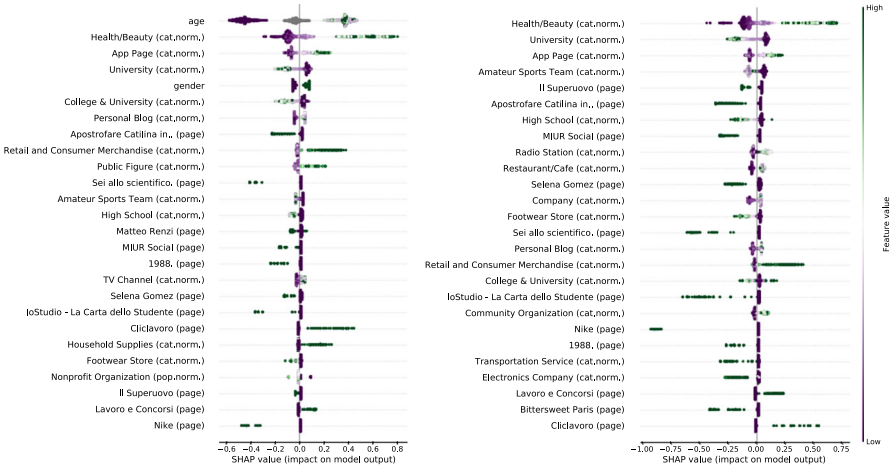

Fig. 2 Feature contributions (via SHAP values). The higher the SHAP value (x-axis) the more the feature contributes to the occupation prediction (unemployment status) in a specific sample. Note that green colors represent higher feature values (Color figure online)

Moving down to the contributions list in Fig. 2, we can see interesting insights emerge. Work-related pages such as *Lavoro e Concorsi* ("work and public service exams", in Italian) and *Cliclavoro* (popular job search portal) are among the most important predictors providing evidence that the models' reasoning is indeed making decisions based on work-related features. Unemployed users also tend to be more interested in App pages as well as retail and home supply pages.

For the employed population, we notice that two trends emerge; one related to the university and schooling and the other on brand and spare time activity pages. For instance, 'University (cat.norm)', 'Io studio' (aka. I study), and 'Sei allo scientifico'(aka. Studying in the Italian technology-oriented High-School) pages are expected since - due to the initial administration of our app - we have an over-representation of students in our dataset which of course are falling under the "employed" label. Spare time activity pages like sports (*Amateur sports team*), travelling (e.g., *1988*), and satyric content *Il Superuovo* (an italian blogger with over half a million followers), and NGO related pages are indicative of employed individuals. Some brand pages were also found as indicative of employment (e.g. *Nike* and *Bittersweet Paris*) in line with findings by Bento et al. (2018) regarding brand engagement and employment in Facebook. The above insights are of particular importance since the adaptive threshold approach does not alter the predictors; hence, the models **Demo+AT.** and **NoDemo+AT.** have the same predictors as the respective simple models.

Fairness. To ensure that the algorithm will make the fairest decisions possible with respect to our basic demographic attributes, we opted for an adaptive threshold applied to the internal decision-making step of the LightGBM classifier (see Method section). Figure 3 illustrates the interplay among fairness, precision, and recall, when altering

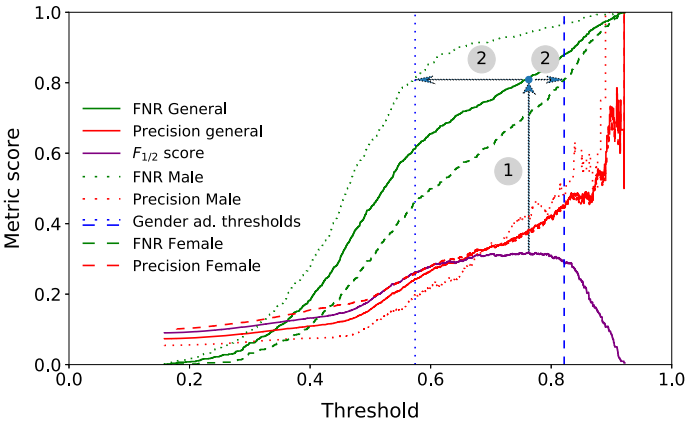

Fig. 3 Fairness threshold optimisation. Curves show the precision (red) and false negative rate (FNR, in green) scores as a function of threshold for the general, male and female population (solid, dotted, and dashed lines, respectively). Blue lines represent the gender thresholds that give the same FNR as the one that maximizes the $F_{1/2}$ score. The numbered points indicated in the graph represent: (1) our starting point, that is, the threshold value that maximises the $F_{1/2}$ score. For this point, we estimate the FNR. Then, (2) we estimate the thresholds that produce that same FNR value for both genders (in the example, .57 for male and .82 for female) (Color figure online)

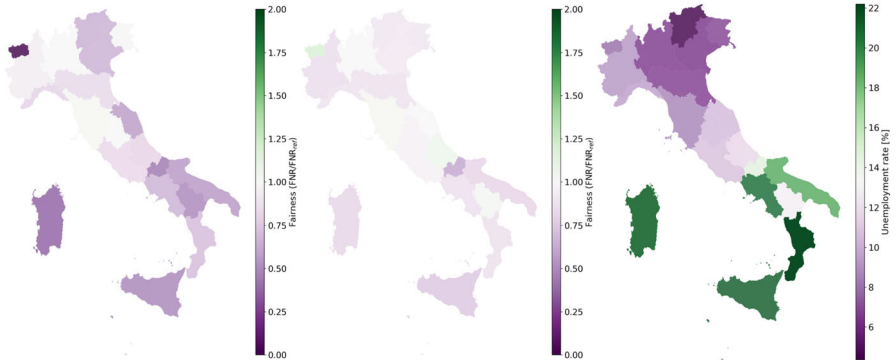

Fig. 4 **a** bf b Regional fairness of the models. Fairness is computed as the FNR in each region, relative to the FNR of the Lombardy reference region. The color extremities are both unfair. **c** Unemployment level per region (Color figure online)

the internal decision threshold of the model. The purple line represents the F_β score for $\beta = 1/2$. We chose this value as it prioritises precision over recall (Baeza-Yates and Ribeiro-Neto 1999), according to our depicted scenario of a cost-efficient campaign. At its maximum (see step (1) in the figure) the general FNR is 0.79, so in step (2) we set the gender thresholds that get this FNR for each gender. These thresholds are depicted in dark blue lines for males and females, respectively.

This approach aims at providing equal opportunities to people from both gender groups while maintaining the accuracy of prediction. As seen in Table 3, the respective

AUC score for the **Demo+AT** and **NoDemo+AT** models naturally remains invariant with respect to the models without the adaptive threshold, as well as the accuracy of each demographic category. However, by satisfying our fairness criterion while prioritizing precision, a trade-off with recall was inevitable.

Table 3 reports the fairness metric scores obtained by each model. Notice that fairness is always reported with respect to a reference group, as described by Equation (1). Hence, we only report the scores for the remaining classes. The disparity intolerance percentage is empirically set to $\tau = 0.80$. If the fairness metric for a group with respect to the reference group is within the $(\tau, \frac{1}{\tau})$ range, then the model is said to be fair to this specific group (see fairness scores in bold in Table 3). The fairness scores of the **Demo** and **NoDemo** models are both very biased towards women since they have approximately half the false negative rates of the males (i.e. the reference group); this bias almost vanishes in the models with adaptive threshold. At the same time, the FNR metric is also improved for all age categories in the **NoDemo+AT.** model, which is the fairest model with respect to gender and all but one age groups.

Figure 4(c) depicts the unemployment rate in Italy per region as reported in the official statistics. Green areas represent regions with higher unemployment rates, while purple areas represent regions with less unemployment. Interestingly, the basic model **NoDemo** is often favouring some regions while disfavouring others, as seen in Fig. 4(a). Both extremities of the scale signal that we are missing people from certain regions disproportionately. More specifically, when values are in deep purple in Fig. 4(a), it implies that we privilege the unemployed in those regions (they are more represented by our predictive model than in the reference region of Lombardy). Thus, ideal fairness values are around 1.0 and have light colour. Hence, the more intense the colour, the more intense the discrimination, while the more transparent the colour, the fairer the model.

When looking into the gender biases of the predictions at a regional level, we notice that the models without adaptive threshold have strong gender biases in almost every region (see Fig. 5(a)), which fade out when the adaptive threshold is applied (see Fig. 5(b)). This is of major importance since we achieve age and geographic fairness while aiming for equal gender representation within the unemployed community. The remaining gender biases in the Sardinia region may be due to uneven population representation in some areas combined with the small population size in our sample.

Generalisability and Limitations. The choice of β depends on the requirements of the specific application. To demonstrate the generalisability of our approach, in the Supplementary Materials we showcase an alternative hypothetical scenario where recall was privileged by design $\beta = 2$.

Another implementation alternative regards the choice of the fairness criterion. Chouldechova (2017) explored the trade-off between fairness measures and, for a binary classification task where the target prevalence is different across groups –like in our case– they showed that a calibrated algorithm cannot guarantee equal false negative and false positive rates in all of them, and one score must be privileged. To guarantee equality of opportunity across groups, we opted for equal FNR. However, this criterion should be modified according to the task; for instance, when interventions have a punitive character equality of opportunity should be replaced by equal false positive rates (FPR) (Saleiro et al. 2018).

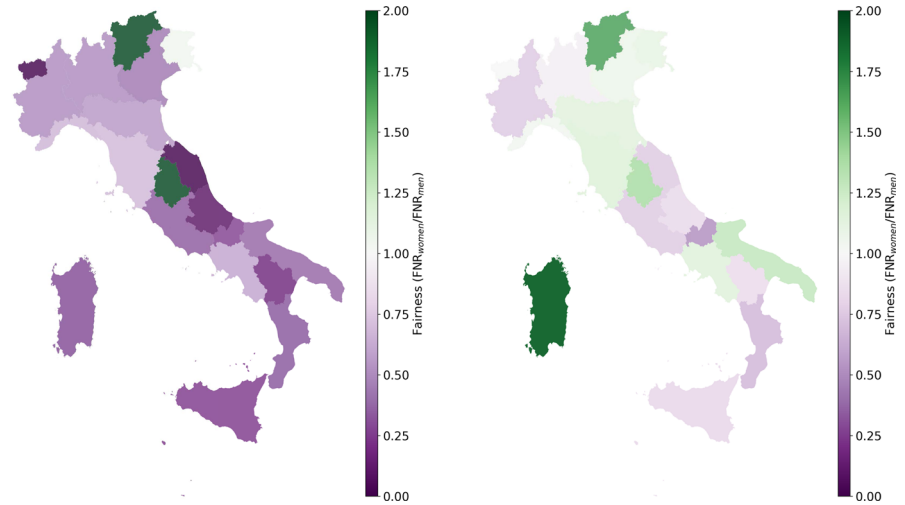

Fig. 5 Gender fairness per region in the NoDemo (left) and NoDemo+Thresh. (right) models. Gender fairness is computed as the FNR of females in relation to that of males. The color extremities are both unfair (Color figure online)

Similar to the related literature (Dutta et al. 2020; Hardt et al. 2016), a limitation of our approach is the requirement for the sensitive attributes at the individual level to be accessible while training the model, for the estimation of the corresponding thresholds. At this point, we need to clarify that once the thresholds are set, these attributes are not required for the further deployment of the models. Alternatives to this have been proposed in the literature employing adversarial learning (Zhang et al. 2018) or regularization (Beutel et al. 2019).

6 Conclusions

Traditionally, researchers and practitioners employed basic demographic information like age, gender, and geography (Goyat 2011; Wood et al. 2019) to gain insights on the population under investigation or maximise the efficiency of their communication campaigns. Today, several socioeconomic attributes and behaviours are assessed by machine learning predictive models to drive actions.

Moreover, AI-driven decision making is increasingly more considered both in policy making and in humanitarian crisis management (Aiken et al. 2022), with the ML models to compete for the highest prediction accuracy often as the sole metric of performance, neglecting to account for fairness metrics. A common ground for all data science for social good predictive tasks is that the vulnerable populations of interest – in our case, the unemployed– consist of significantly fewer samples with respect to the majority class. Often the focus is on developing predictive models that outperform the

state of the art accuracy score, overlooking whether these models introduce or amplify existing discrimination in society.

To achieve "fairness" a solution is to invest in better and more data; however, this is almost never possible in practice. Here, we propose a simple and efficient approach that provides fair classifications regardless of the demographic attribute of choice while the method can be extended to a combination of attributes.

The study presented here was specifically designed to address a not-for-profit organisation's need, whose aim was to communicate educational and job opportunities within the younger unemployed community. To this end, we designed and developed an ad-hoc Facebook-hosted application, reaching out to approximately 64k participants. To automatically predict the occupational status of these users, we postulated the study as a supervised classification task. Our models, inferring solely on users' "Likes" on Facebook Pages, were able to predict the occupational status (employed vs unemployed) with an AUC of .74. Despite achieving a satisfactory accuracy in prediction, we dived further into the demographic breakdown of the obtained insights, discovering biases in both gender and age attributes. We showed that the straightforward solution of "hiding" those sensitive information from the model does not ensure a fair prediction as they are often embedded not only in other demographic but also in our digital behavioural patterns and are likely to influence the models' decision.

Hence, we proposed an approach based on adaptive thresholding of the predictive model's decision-making step. This method ensures that the model makes as fair predictions as possible, according to the most adequate fairness metric for the task, which in our specific case is the parity of opportunity (FNR), or else ensuring that the model will not disproportionately miss individuals from specific protected groups. Interestingly, minor inequalities (i.e., geographical) are likely to improve when assessing the substantial ones (i.e. gender). Our framework includes a double out-of-sample evaluation providing stability in unseen elements for real-life applications.

This simple modification can be directly applied to assistive interventions that rely on AI for their communication campaigns or policy design to provide fair and explainable results. Most importantly, the method is flexible, both, to be extended in other demographic features, and, to other scenarios and concepts of "fairness". Preserving the privacy of potentially vulnerable populations is essential since increasingly more new data sources are employed to complement traditional methods for targeting humanitarian assistance, particularly in crisis settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10618-022-00855-y.

Acknowledgements K.K acknowledges support from the "Lagrange Project" of the ISI Foundation funded by the Fondazione CRT.

References

Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H (2018) A reductions approach to fair classification. In: International Conference on Machine Learning, pp 60–69. PMLR

Aiken E, Bellue S, Karlan D, Udry C, Blumenstock JE (2022) Machine learning and phone data can improve targeting of humanitarian aid. Nature 1–7
Akintande OJ (2021) Algorithm fairness through data inclusion, participation, and reciprocity. In: International Conference on Database Systems for Advanced Applications, Springer, pp 633–637
Baeza-Yates R, Ribeiro-Neto B et al (1999) Modern Information Retrieval, vol 463. ACM Press, New York
Barocas S, Selbst AD (2016) Big data's disparate impact. Calif L Rev 104:671
Becker GS (2010) The Economics of Discrimination. University of Chicago Press, Chicago
Bento M, Martinez LM, Martinez LF (2018) Brand engagement and search for brands on social media: Comparing generations x and y in portugal. J of Retailing and Consum Serv 43:234–241
Beutel A, Chen J, Doshi T, Qian H, Woodruff A, Luu C, Kreitmann P, Bischof J, Chi EH (2019) Putting fairness principles into practice: Challenges, metrics, and improvements. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp 453–459
Bi B, Shokouhi M, Kosinski M, Graepel T (2013) Inferring the demographics of search users: Social data meets search queries. In: Proceedings of the 22Nd International Conference on World Wide Web. WWW '13, ACM, New York, NY, USA, pp 131–140. https://doi.org/10.1145/2488388.2488401
Bokányi E, Lábszki Z, Vattay G (2017) Prediction of employment and unemployment rates from twitter daily rhythms in the us. EPJ Data Sci 6(1):14
Bonanomi A, Rosina A, Cattuto C, Kalimeri K (2017) Understanding youth unemployment in italy via social media data. In: 28th IUSSP International Population Conference, Cape Town, South Africa
Calders T, Verwer S (2010) Three naive bayes approaches for discrimination-free classification. Data mining and knowl discov 21(2):277–292
Chhabra A, Masalkovaitė K, Mohapatra P (2021) An overview of fairness in clustering. IEEE Access
Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data 5(2):153–163
Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17, Association for Computing Machinery, New York, NY, USA pp 797–806. https://doi.org/10.1145/3097983.3098095
Desiere S, Langenbucher K, et al. (2018) Profiling tools for early identification of jobseekers who need extra support. OECD Policy Brief on Activation Policies (dec) 1–4
Desiere S, Struyven L (2020) Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off. Journal Of Social Policy
Dong Y, Yang Y, Tang J, Yang Y, Chawla NV (2014) Inferring user demographics and social strategies in mobile social networks. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, USA, pp 15–24. https://doi.org/10.1145/2623330.2623703
Dutta S, Wei D, Yueksel H, Chen P-Y, Liu S, Varshney K (2020) Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In: International Conference on Machine Learning, pp 2803–2813. PMLR
Eslami, M., Krishna Kumaran, S.R., Sandvig, C., Karahalios, K.: Communicating algorithmic process in online behavioral advertising. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2018)
Fatehkia M, Kashyap R, Weber I (2018) Using facebook ad data to track the global digital gender gap. World Dev 107:189–209
Fatehkia M, Coles B, Ofli F, Weber I (2020) The relative value of facebook advertising data for poverty mapping. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp 934–938
Felbo B, Sundsøy P, Lehmann S, de Montjoye Y-A et al. (2017) Modeling the temporal nature of human behavior for demographics prediction. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 140–152
Gao J, Zhang Y-C, Zhou T (2019) Computational socioeconomics. Physics Reports
Goel S, Hofman J, Sirer MI (2012) Who does what on the web: Studying web browsing behavior at scale. In: International Conference on Weblogs and Social Media, pp 130–137
Goyat S (2011) The basis of market segmentation: A critical review of literature. Eur J of Bus and Management 3(9):45–54

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16, Red Hook, NY, USA, pp 3323–3331

ISTAT (2020) ISTAT Database. Data on unemployed rate. http://dati.istat.it

Kalimeri K, Beiró MG, Delfino M, Raleigh R, Cattuto C (2019) Predicting demographics, moral foundations, and human values from digital behaviours. Comput in Human Behav 92:428–445

Kalimeri K, Beiró MG, Bonanomi A, Rosina A, Cattuto C (2020) Traditional versus facebook-based surveys: Evaluation of biases in self-reported demographic and psychometric information. Demogr Res 42(5):133–148

Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl and Inf Syst 33(1):1–33

Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 35–50

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, pp 3146–3154

Kilbertus N, Rojas Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B (2017) Avoiding discrimination through causal reasoning. Advances in neural information processing systems 30

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807

Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc of the National Acad of Sci 110(15):5802–5805

Kuhn P (1987) Sex discrimination in labor markets: The role of statistical evidence. The American Economic Review 567–583

Leonelli S, Lovell R, Wheeler BW, Fleming L, Williams H (2021) From fair data to fair data use: Methodological data fairness in health-related social media research. Big Data & Soc 8(1):20539517211010310

Llorente A, Garcia-Herranz M, Cebrian M, Moro E (2015) Social media fingerprints of unemployment. PLOS ONE 10(5):1–13

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2019) Explainable AI for Trees: From Local Explanations to Global Understanding

Lundberg SM, Lee S-I (2017a) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30, pp 4765–4774

Lundberg S, Lee S-I (2017b) A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874

Malmi E, Weber I (2016) You are what apps you use: Demographic prediction based on user's apps. ICWSM, 635–638

Mason SJ, Graham NE (2002) Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. Quarterly J of the Royal Meteorol Soc 128(584):2145–2166

Matz SC, Menges JI, Stillwell DJ, Schwartz HA (2019) Predicting individual-level income from facebook profiles. PloS one 14(3):0214369

Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M-E, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E et al (2020) Bias in data-driven artificial intelligence systems-an introductory survey. Wiley Int Rev: Data Mining and Knowl Discov 10(3):1356

Olteanu A, Castillo C, Diaz F, Kıcıman E (2019) Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data 2:13

Olteanu A, Castillo C, Diaz F, Kiciman E (2016) Social data: Biases, methodological pitfalls, and ethical boundaries. https://doi.org/10.2139/ssrn.2886526

O'Neil C (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, New York

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J of Mach Learning Res 12:2825–2830

Pedreshi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 560–568

Pessach D, Shmueli E (2022) A review on fairness in machine learning. ACM Comput Surveys (CSUR) 55(3):1–44
Rama D, Mejova Y, Tizzoni M, Kalimeri K, Weber I (2020) Facebook ads as a demographic tool to measure the urban-rural divide. In: Proceedings of The Web Conference 2020, pp 327–338
Saleiro P, Kuester B, Stevens A, Anisfeld A, Hinkson L, London J, Ghani R (2018) Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577
Seneviratne S, Seneviratne A, Mohapatra P, Mahanti A (2015) Your installed apps reveal your gender and more! ACM SIGMOBILE Mobile Comput and Commun Rev 18(3):55–61
Stoll MA, Raphael S, Holzer HJ (2004) Black job applicants and the hiring officer's race. ILR Rev 57(2):267–287
Sundsøy P, Bjelland J, Reme B-A, Jahani E, Wetter E, Bengtsson L (2016) Estimating individual employment status using mobile phone network data. arXiv preprint arXiv:1612.03870
Toole JL, Lin Y-R, Muehlegger E, Shoag D, González MC, Lazer D (2015) Tracking employment shocks using mobile phone data. J of The Royal Soc Int 12(107):20150185
Urbinati A, Kalimeri K, Bonanomi A, Rosina A, Cattuto C, Paolotti D (2020) Young adult unemployment through the lens of social media: Italy as a case study. In: International Conference on Social Informatics, Springer, Cham, pp 380–396
van Landeghem B, Desiere S, Struyven L (2021) Statistical profiling of unemployed jobseekers. IZA World of Labor, Germany
Van Rossum G, Drake FL (2009) Python 3 Reference Manual. CreateSpace, Scotts Valley, CA
Verma S, Rubin J (2018) Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (fairware), pp 1–7. IEEE
Wood R, Murch B, Betteridge R (2019) A comparison of population segmentation methods. Oper Res for Health Care 22:100192
Yeung K, Lodge M (2019) The Possibilities of Digital Discrimination: Research on E-commerce, Algorithms and Big Data. Oxford University Press, UK
Ying JJ-C, Chang Y-J, Huang C-M, Tseng VS (2012) Demographic prediction based on users mobile behaviors. Mobile Data Challenge
Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web, pp 1171–1180
Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: International Conference on Machine Learning, pp 325–333. PMLR
Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp 335–340
Zhong Y, Yuan NJ, Zhong W, Zhang F, Xie X (2015) You are where you go: Inferring demographic attributes from location check-ins. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. WSDM '15, ACM, New York, NY, USA, pp 295–304