Fairness in Vulnerable Attribute Prediction on Social Media

Keywords: fairness, data for good, policy making, vulnerable populations

Extended Abstract

The plentifulness of social media data combined with artificial intelligence (AI) methods is having a massive impact on interdisciplinary fields such as social sciences and humanities, allowing for actionable insights into populations that were previously hard to reach. All data sources, though, come with their own biases and limitations [1], and since machine learning is heavily based on the statistical properties of data, caution is required to avoid generating or amplifying existing social inequalities. This is particularly important since, in computer science, the performance score would be the sole criterion for optimising a model, which may lead to AI-empowered decisions that treat individuals differently based on their characteristics such as gender, age or other attributes [2].

This study was designed to address the needs of an Italian nonprofit organisation, which aimed at reaching out to young unemployed individuals with educational and job opportunities via communication channels that are more likely to appeal to younger generations. To that extent, an ad-hoc META-hosted app was developed whose main functionality was administering questionnaires while gauging the participants' Likes on META Pages. The final scope of the application would be to propose a job-related opportunity to those classified as "unemployed" by the machine learning (ML) model. 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 beyond creating 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?*

The concept of "fairness" is inherently subjective and can have different interpretations and definitions depending on the specific problem under investigation [3]. 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. To answer these questions, we build a series of ML models, assessing the predictive power of digital and demographic data regarding accuracy (AUC[1]) and fairness. Although our ML models achieve state-of-the-art performance in accurate automatic prediction of the employment status (.74AUC), we dive further into highlighting the biases and trade-offs when introducing the notion of fairness expressed as the equality of opportunity.

Our classification scheme includes a 10-fold cross-validation approach to avoid overfitting while the interpretability of the most predictive features as evidenced by the SHAP (Shapley Additive exPlanations) technique. To ensure fairness, we build on the seminal work of Hardt et al. [4], introducing an adaptive threshold criterion on the default threshold that binary ensemble

[1] We conventionally refer to the AUROC values as "accuracy".

classifiers have as a reference. Figure 1 illustrates the interplay among fairness, precision, and recall when altering the internal decision threshold of the model. We force the threshold to prioritise precision over recall for our task. We also perform experiments excluding the known demographic attributes to assess the effectiveness of the "fairness through unawareness" approach.

Importantly, in line with evidence from the current literature [5, 6], we confirm that excluding protected attributes from the ML model's set of predictors does not, per se, guarantee a "fair" model since sociodemographic attributes might be embedded in our digital behavioural patterns. We also demonstrate the generalisation potentials of this approach in other demographic attributes, while it can also be applied to a combination of attributes. After the application of the adaptive threshold approach, the overall AUC score for the model remained invariant, however, to satisfy our fairness criterion which prioritised precision, a trade-off with the recall was inevitable. Finally, we provide vital observations to help practitioners generalise this approach to other domains, for instance, humanitarian crisis management, where AI models are increasingly more involved in the decision-making process.

References

[1] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman, "Social data: Biases, methodological pitfalls, and ethical boundaries," 2016.

[2] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown, 2016.

[3] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 IEEE/ACM International Workshop on Software Fairness (fairware)*. IEEE, 2018, pp. 1–7.

[4] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16, Red Hook, NY, USA, 2016, p. 3323–3331.

[5] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568.

[6] K. Kalimeri, M. G. Beiró, M. Delfino, R. Raleigh, and C. Cattuto, "Predicting demographics, moral foundations, and human values from digital behaviours," *Computers in Human Behavior*, vol. 92, pp. 428–445, 2019.

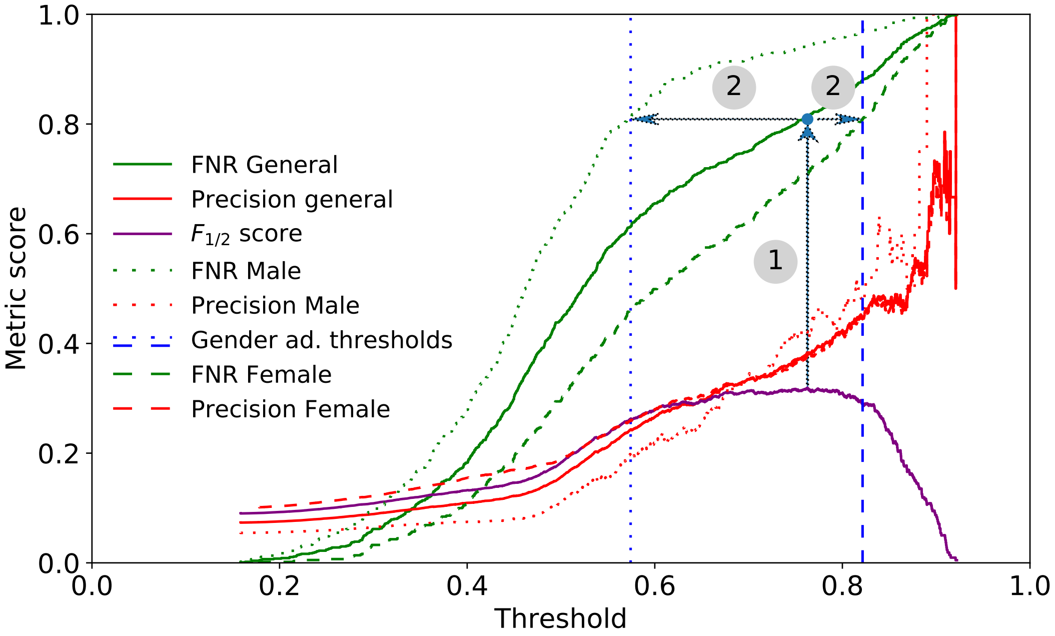

Figure 1: **Fairness threshold optimisation**. Curves show the precision (red) and false negative rate (FNR, in green) scores as a function of threshold for the general, male and female population (solid, dotted, and dashed lines, respectively). Blue lines represent the gender thresholds that give the 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).