

Addressing Bias in LLMs: Strategies and Application to Fair AI-based Recruitment

Alejandro Peña, Julian Fierrez, Aythami Morales, Gonzalo Mancera, Miguel Lopez-Duran,
Ruben Tolosana

Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain
{alejandro.penna, julian.fierrez, aythami.morales, gonzalo.mancera,
miguel.lopezd, ruben.tolosana}@uam.es

Abstract

The use of language technologies in high-stake settings is increasing in recent years, mostly motivated by the success of Large Language Models (LLMs). However, despite the great performance of LLMs, they are susceptible to ethical concerns, such as demographic biases, accountability, or privacy. This work seeks to analyze the capacity of Transformers-based systems to learn demographic biases present in the data, using a case study on AI-based automated recruitment. We propose a privacy-enhancing framework to reduce gender information from the learning pipeline as a way to mitigate biased behaviors in the final tools. Our experiments analyze the influence of data biases on systems built on two different LLMs, and how the proposed framework effectively prevents trained systems from reproducing the bias in the data.

1 Introduction

During the last decade, automatic decision-making systems have increased their presence in society, as AI-based technologies have continued to improve their performance in a vast diversity of tasks. A 2016 report from the White House (Executive Office of the President 2016) highlighted the key role that big data and automatic systems would play in high-stake domains such as employment (Peña et al. 2023), education (Daza et al. 2023; Becerra et al. 2025), or lending access. Since then, major developments have taken place in the field of artificial intelligence (AI). Specifically in the Natural Language Processing (NLP) domain, the Transformers model, based on self-attention mechanisms (Vaswani et al. 2017), supposed a paradigm shift with respect to previous recurrence-based models and paved the way for great models to come. Today, Large Language Models (LLMs) and foundational models such as novel GPT models (OpenAI 2023), LLaMA (Touvron et al. 2023), or DeepSeek, along with techniques that facilitate their fine-tuning to downstream tasks (Hu et al. 2021), enable unprecedented applications of this technology.

Although NLP systems have proven to be an extremely useful tool for different applications (Peña et al. 2023), concerns remain about how to properly use this technology to prevent harmful consequences. In this sense, in December 2023 the European Parliament and the Council reached an

agreement on the new AI Act, which establishes a risk-based categorization of AI systems, setting different requirements for each level. The final objective is to ensure the fundamental rights of EU citizens by creating an AI environment in which automatic systems deployed in Europe comply with the basic principles of a socially responsible AI (Cheng, Varshney, and Liu 2021), such as privacy (Hassanpour et al. 2022), or demographic fairness (Caliskan, Bryson, and Narayanan 2017; Blodgett et al. 2020; Serna et al. 2022b). The recent emergence of foundational models and generative AI technologies has even highlighted the urgency of new regulations to address challenges related to these technologies (Bommasani et al. 2023), with the final objective of ensuring citizens' rights. In the US, a debate around the need to create new AI regulatory agencies is also taking place, with a focus on the concerns exposed.

In terms of the labor market and employment, the previously mentioned advances in AI at large have long contributed to this domain (e.g., AI-based personalized recommendations in employment platforms such as LinkedIn). Automatic decision systems have been of great use in different stages of the recruitment pipeline (Singla et al. 2024; Bogen and Rieke 2018). Recent advances in prompt-based LLMs are encouraging the development of new tools¹ to help recruiters find the best suitable candidates for a job. As stated in the 2016 White House report, the use of AI in employment presents some opportunities, for example, the prevention of human affinity biases during the screening or scoring of candidates. However, due to historical biases in the labor market (Bertrand and Mullainathan 2004; Bendick Jr, Jackson, and Romero 1997; Altonji and Blank 1999), a careful design of these systems with fairness perspectives is required (Schumann et al. 2020; Sánchez, Dençik, and Edwards 2020). Some algorithms have exhibited biases in this domain (Dastin 2018), and therefore solutions to bias-free procedures are being proposed. These solutions range from addressing fairness in labor-related technologies, such as automatic ranking algorithms (Yang et al. 2017; Zehlike et al. 2017) or screening processes (Haryan et al. 2024; Mancera et al. 2025b), to projects that focus on preventing discrimination in algorithmic hiring.²

¹<https://www.herohunt.ai/blog/how-ais-like-gpt-are-changing-recruitment>

²<https://www.findhr.eu/>

In this work, we address the topic of demographic biases in Transformer-based technologies, with application to the case study of automated recruitment tools, which are considered high-risk systems by Regulations like the EU AI Act. With the aim of preventing data biases from affecting such systems, we present a privacy-enhancing learning framework that reduces gender information in the learning pipeline. In particular, we propose and evaluate two different methods to reduce demographic information in the system. One of these methods relies on deep model explainability, while the other is framed within the state of the art in bias mitigation in the image domain (Serna et al. 2021, 2022a; DeAlcala et al. 2023). We analyze the bias learned by a recruitment tool based on two different Transformers models, i.e. BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019), and adapt the proposed methods to reduce gender information in the learning process.

2 Fairness Research in NLP

The subject of fairness and algorithmic discrimination has been widely studied in the area of biometrics research (Acien et al. 2018; Terhorst et al. 2022; Serna, Morales, and Fierrez 2025), where some popular technology benchmarks include specific tasks dealing with bias assessment (Stragapede et al. 2024; Melzi et al. 2024). Fairness research has also attracted great interest in NLP in recent years, following the general trend to achieve responsible systems in the AI community (Cheng, Varshney, and Liu 2021). However, as some authors have pointed out (Blodgett et al. 2020), most of the work that analyzes bias in NLP fails to properly define which behavior of the system is harmful, what is meant by a biased system (Serna et al. 2020), or the source of these biases. Several works have addressed gender bias in language technologies, due to the societal implications and prevalence of this kind of discrimination. For example, a great deal of attention has been paid to gender bias in language representations such as word embeddings (Bolukbasi et al. 2016; Chen et al. 2021; Garg et al. 2018). De-Arteaga et al. (2019) analyzed different semantic representations and their relationship with the gender bias exhibited by occupation classifiers trained in Common Crawl biographies. They found strong gender biases that reflect historical biases in the labor domain using all three representations. Both Caliskan, Bryson, and Narayanan (2017) and Bolukbasi et al. (2016) proposed bias metrics to analyze word embeddings, which they applied to the same pre-trained embeddings simultaneously, i.e. GloVe embeddings, arriving to similar conclusions. In addition to gender bias, Caliskan, Bryson, and Narayanan (2017) measured human-like biases in word embeddings, finding that these captured commonly human prejudices well documented in psychology. As showed by (Garg et al. 2018), word embeddings trained on 100 years of text data quantify historical demographic stereotypes, such as occupation rates. Social biases were also found in Transformers-based encoder representations (May et al. 2019), including intersectional biases such as the black woman stereotype.

Going beyond stereotypical associations in semantic representations, different works have analyzed demographic bi-

ases in various NLP tasks, including natural language inference (Rudinger, May, and Van Durme 2017), language generation (Sheng et al. 2019), sentiment analysis (Kiritchenko and Mohammad 2018), hate speech detection (Sap et al. 2019; Dixon et al. 2018; Park, Shin, and Fung 2018), machine translation (Stanovsky, Smith, and Zettlemoyer 2019; Prates, Avelar, and Lamb 2020), or co-reference resolution (Zhao et al. 2018; Cao and Daumé 2021), among others. As we aforementioned, different authors have pointed out the social perspectives and stereotypes depicted in text datasets as the source of these biases. A good example of this relationship can be found in the work analyzing social biases in toxicity and the detection of hate speech, where identity terms associated with certain demographic groups that are the target of hateful speech itself, such as women (Park, Shin, and Fung 2018) or queer people (Dixon et al. 2018), are overrepresented in toxic samples. An unbalanced representation of identity terms is not the only way to introduce bias in a dataset, as noted in (Sap et al. 2019), where racial disparities were found to be a consequence of annotation biases due to the use of the African-American English dialect. It should be noted that some work has proposed methods to mitigate the biases learned by NLP models, such as removing sensitive indicators from input text (De-Arteaga et al. 2019), reducing the correlation between model output and embeddings of words that represent sensitive words (Romanov et al. 2019), data augmentation techniques (Qian et al. 2022), or even relying on adversarial prompting and safe response generation (Ge et al. 2023). Some work has relied on improving demographic privacy as a bias mitigation approach, for example, using NER technologies to anonymize input data (Mancera et al. 2025b).

As argued by Blodgett *et al.*, one recurring problem in NLP works dealing with fairness is the vague description of how they conceptualize bias and how the behavior of the system (Rahwan et al. 2019; Serna, Morales, and Fierrez 2025) can be harmful (Serna et al. 2020) and to whom (Blodgett et al. 2020). For example, Caliskan *et al.*, argued that the codification of a prejudice (that is, what can be considered as a bias) is not harmful *per se*, but it is the application context that determines whether that behavior would be harmful (Caliskan, Bryson, and Narayanan 2017). In general, we can categorize harmful system behavior into three different categories, namely: *i) Representational harm*, which occurs when the system represents certain groups in a less favorable manner, or directly erases them; *ii) allocational harm*, when the system unfairly assigns resources differently across groups; and *iii) quality of service harm*, which can occur when a system does not work the same for different groups. Although the varying performance of a sentiment analysis model depending on the demographics in the data samples could be a clear example of the latter (Kiritchenko and Mohammad 2018), the prejudices encoded in word embedding representations as we aforementioned can suppose representational harm in situations where these stereotypical associations are reproduced. Our work focuses on a case of allocational harm, where the demographic bias learned by the system can harm certain groups by denying the opportunities of those individuals to access a job.

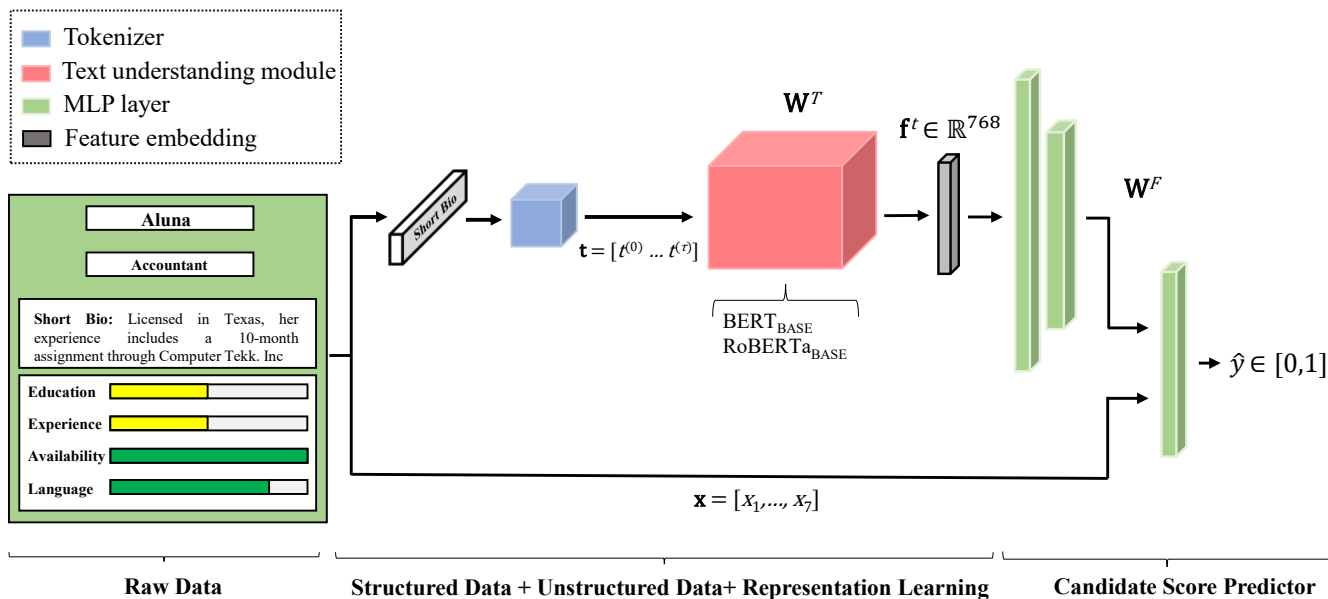


Figure 1: Learning architecture for the resume scoring tool \mathbf{W}^S . The system receives a multimodal input resume, composed by a candidate competencies vector \mathbf{x} and a biography \mathbf{t} . A text understanding module \mathbf{W}^T processes the bios to obtain a feature vector \mathbf{f}^t . Then, both \mathbf{f}^t and \mathbf{x} are fed to a fusion module \mathbf{W}^F to obtain the final score \hat{y} .

3 Problem Formulation

The aim is to train a fair scoring tool \mathbf{W}^S , based on structured and non-structured textual information. Concretely, suppose a multimodal dataset of resumes \mathcal{D} , where the i -th resume is composed of a vector of attributes or candidate competencies \mathbf{x}_i , and a biography, represented by a sequence of τ tokens $\mathbf{t}_i = [t_i^{(1)}, \dots, t_i^{(\tau)}]$ after tokenization. The resumes have a corresponding score $y \in [0, 1]$, assigned by human annotators. Furthermore, we have access to the gender of the applicant i , which is a binary attribute $z_i \in \{0, 1\}$ (0 for men, 1 for women).³ In this context, the scoring tool \mathbf{W}^S could be built on the general learning architecture of Fig. 1. This system is trained in \mathcal{D} to correctly predict a new score \hat{y}_i that approximates y_i based on an input resume $\{\mathbf{x}_i, \mathbf{t}_i\}$. Note that the scoring tool is composed of: *i*) a text understanding module \mathbf{W}^T , which takes a tokenized biography as input and outputs a text embedding $\mathbf{f}^t \in \mathbb{R}^N$; and *ii*) a fusion module \mathbf{W}^F , which further processes \mathbf{f}^t , and merges it with competencies \mathbf{x} to predict a score \hat{y} .

Suppose now that the scores are not a completely objective assessment of the candidate but a certain degree of sub-

³Although we adopt a gender definition that aligns in some points (e.g., gender as a binary attribute) with traditional definitions of such (Keyes 2018), our intention is to improve the research on fairness in AI, for which we rely on existing data with such limitations, while considering the research conducted here to be extrapolable to other definitions. As some authors have pointed out, a crucial step in preventing social injustice and demographic bias in AI is to properly define these terms and consider the perspectives of individuals belonging to these groups (Leavy, Siapera, and O’Sullivan 2021), for which a change to new inclusive definitions is needed (Cao and Daumé 2021).

jectivity was included during their generation. Without loss of generality, suppose that human annotators assigned lower scores to women compared to men. Subjectivity during annotation leads to a new set of gender-biased scores y^G . In this context, a data-driven learning framework could exploit the demographic information in the biographies and use it to learn bias in scores y^G , leading to a scoring tool that unintentionally discriminates by gender. Using these scores to select potential candidates for a job, for example, by providing a short list of the best suitable applicants, would result in a case of allocational harm (Blodgett et al. 2020), as the scores \hat{y} calculated by the tool would unfairly assign better opportunities to one demographic group using information about the sensitive attribute z (i.e., under the same conditions, the score for men would be higher than the score for women). In order to prevent this behavior from the recruitment tool to unfairly discriminate again a demographic groups, we explore here the application of a privacy-enhancing framework as a bias mitigation strategy.

4 Bias Reduction: Proposed Methods

We propose to remove gender information from the previously introduced learning framework with two different approaches (see Fig. 2). Each method focuses on a different stage of the learning process. The idea is that, by removing gender information, the network will not be able to establish a relationship between the sensitive attribute z (gender in this case) and the bias in the scores, therefore preventing it from reproducing the bias.

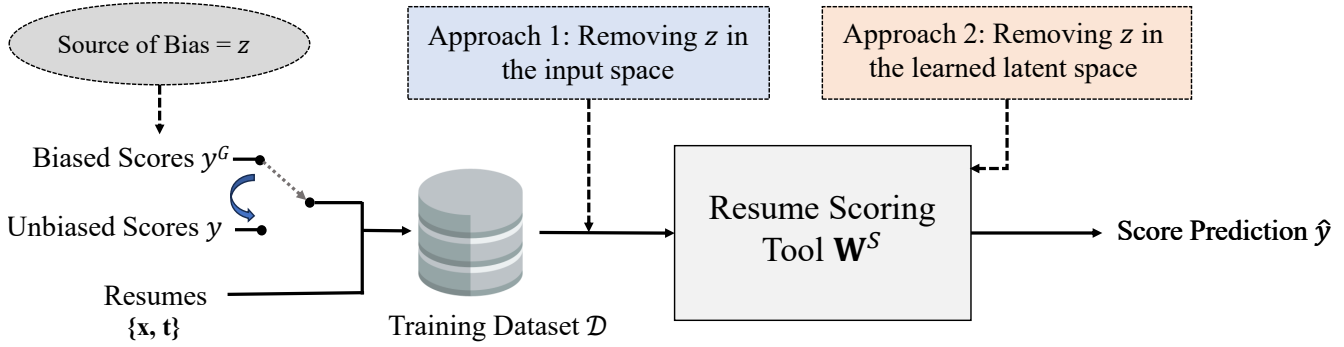


Figure 2: Proposed privacy-enhancing learning framework for bias mitigation. We aim to mitigate the effect of demographic data biases by removing demographic information in the input space or in latent space of an AI-based resume scoring tool.

4.1 Via Model Explainability

This approach performs an iterative learning process in which potential biased tokens are detected and removed from the input sentences in each iteration, followed by a network retraining. This could be achieved by applying the Integrated Gradients technique to the model (Sundararajan, Taly, and Yan 2017). Integrated Gradients (IG) is a model explainability technique that allows us to determine the importance of input features in the model’s predictions, without requiring us to modify the model. This method begins with a baseline input f' , which could be a zero vector, and builds a sequence of data samples as the interpolation between f' and the input at hand f_i . For each data sample, the gradients are computed to measure how changes in the features affect the model’s prediction. Gradients are accumulated, resulting in an attribution mask that resembles the relevance of the characteristic.

This technique measures token relevance in a validation set and detects relevant words in the prediction that do not appear to correlate with the target task or that explicitly correlate with the sensitive attribute. Our hypothesis is that, in the presence of data biases that affect the predictions of the model, some tokens will act as proxies for the sensitive attribute z , thus appearing as relevant tokens for the predictions. This approach improves the process in a data-driven way with respect to De-Arteaga et al. (2019), where, after removing gendered pronouns and names from biographies, the authors concluded that some tokens acted as proxies for the gender attribute. Thus, measuring token relevance in a biased prediction context could be useful for detecting words that act as such proxies but have little impact on the main task.

4.2 Via Adversarial Learning

Our second approach is based on domain-adversarial learning for image analysis applications. We propose adapting the method of (Kim et al. 2019), namely learning-not-to-learn (LNTL), to remove gender information from hidden representations of the scoring tool W^S . Originally proposed to prevent a network from learning a data bias unrelated to the main task during training, the method can be applied as a

privacy enhancing technique to remove sensitive information from latent representations (Peña et al. 2021; Morales et al. 2021). This method proposes to reduce the mutual information between the hidden representation of a network and the data bias through the following loss function that we adapt to our case:

$$\mathcal{L}_{\text{RMSE}}(\{\mathbf{x}, \mathbf{t}\}, y | \mathbf{W}^S) + \lambda \mathcal{I}(z; f(\{\mathbf{x}, \mathbf{t}\} | \mathbf{W}^S)) \quad (1)$$

where $\mathcal{L}_{\text{RMSE}}$ is the Root Mean Squared Error Loss (i.e., we consider the score prediction task as a regression problem), \mathcal{I} is mutual information, and $f(\{\mathbf{x}, \mathbf{t}\} | \mathbf{W}^S)$ is a latent hidden representation of W^S , from which we want to remove the information. The mutual information in Eq. 1 is reduced by optimizing the negative conditional entropy $-\mathcal{H}(z | f(\{\mathbf{x}, \mathbf{t}\}))$, for which the *a posteriori* distribution is required. The authors of LNTL proposed to attach an auxiliary classifier to the network in order to predict the bias (i.e. z here) from the latent representations and use its output to approximate such distribution. In the context of the learning architecture of Fig. 1, the classifier would be attached to the output of an intermediate layer of W^F . To train that intermediate classifier, here a min-max adversarial learning framework (Goodfellow et al. 2014) is used, where we are simultaneously learning the main task (i.e. score prediction), while using an auxiliary prediction network to reduce gender information from latent representations of W^F . Consequently, the characteristics extracted by W^S make it increasingly difficult for the auxiliary classifier to predict z as training progresses, that is, our original objective.

5 Materials and Methods

5.1 Dataset: FairCVdb

FairCVdb is a database for multimodal fairness research in AI (Peña et al. 2020a), with which various fairness testbeds have recently been defined (Peña et al. 2020b, 2023). The database comprises 24,000 synthetic profiles, which were generated by using the identities contained in the DiveFace Face Recognition database (Morales et al. 2021) and the biographies of (De-Arteaga et al. 2019). Each profile includes the following information, usually present in a resume:

- **Non-structured Textual Information:** Each sample comprises two short biographies (a raw-text bio and a gender-agnostic bio), an occupation, and a name. FairCV spans 10 different occupations from 4 different sectors.
- **Structured Information:** FairCV includes 7 features from the 5 information blocks, including education, availability, previous experience, recommendation, and language proficiency. These features are known as candidate competencies. In addition to the competencies, gender is also available.

Each resume was scored using a weighted linear combination of the candidate’s competencies and the labor sector (weights defined by experts on human resources). Although these scores were generated without any consideration of demographic information, thus becoming blind scores, the authors generated two additional sets of scores that included demographic biases, i.e., gender- and ethnicity-biased scores. For the scope of this work, we are only interested in using blind and gender-biased scores. Furthermore, we consider the blind scores as the ground truth.

5.2 Language Models

We will employ two different language understanding models to process the biographies of FairCVdb:

- **BERT** (Devlin et al. 2019): The BERT model was the first transformed-based language model (Vaswani et al. 2017) to apply bidirectional training to learn language representations. Thus, BERT is able to attend to both the left and right context in the sentence (i.e., bidirectional self-attention), and process the whole input at once. In order to achieve this, BERT employs a Masked-Language Modeling (MLM) pre-training objective, where arbitrary tokens are masked, and the network has to predict them based on the surrounding context. In addition to the MLM objective, they include a Next-Sentence Prediction (NSP) loss to improve the model’s ability to understand the relationship between two different sentences. Two versions of BERT are available, the base version (12 layers, with 12 attention heads and a hidden size of 768 features) and the large version (24 layers, with 16 attention heads, and hidden size of 1024 features).
- **RoBERTa:** The RoBERTa model is a variant of the BERT model. (Liu et al. 2019) conducted a study on the BERT pre-training strategy, in which they found that it was indeed undertrained. They used the same architecture, and pre-trained the model in a larger text corpora, carefully selecting the hyperparameters (e.g., a larger batch size). Another key finding of such work was the performance improvement in the downstream task by removing the NSP loss from the pre-training objective. Consequently, the model was pre-trained using only MLM loss. Whilst in BERT’s implementation for the MLM loss tokens were randomly masked once during preprocessing, RoBERTa’s training strategy allows us to randomly mask tokens in each iteration, hence obtaining a much larger corpus of potential masking combinations. Similarly to BERT, RoBERTa is available in the same base and in large versions.

5.3 Fairness Criteria

In this work, we train a fair resume scoring tool in the presence of gender biases (originated due to annotation bias in the dataset). When implemented, the resume scoring system would assign a predicted score \hat{y} to applicant resumes to help human recruiters find the most suitable candidates for a job. One way to do this is by computing a shortlist of the best suitable candidates, sorted by \hat{y} . Note that while the scoring tool itself is a regression model (i.e., it predicts a continuous score $\hat{y} \in [0, 1]$), the recruitment process can be modeled as a binary classification problem, in which the predicted score \hat{y} is the confidence used to determine whether the candidate appears on such a shortlist. In order to assess the (un)fairness of the whole process resulting from the predictions \hat{y} , we use two fairness metrics commonly applied in traditional Machine Learning (Hardt, Price, and Srebro 2016):

- **Statistical Parity.** This criterion requires the probability that a decision is independent of membership in sensitive groups. In our case, this leads to an equal probability of appearing in the shortlist across genders, which can be approximated by the gender proportions within it. Failing to achieve this criterion suppose an *evidence* of allocational harm, but it does not conclude so in general since it does not consider features relevant to the decision.

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1) \quad (2)$$

- **Equality of Opportunity.** This criterion is commonly used in classification tasks in which the positive outcome is associated with an advantaged decision. The criterion ensures that the automatic process performs equally well across sensitive groups. To do so, it requires independence of the sensitive attribute subject to a positive ground truth (i.e. similar recall), measuring only whether individuals that deserve the positive outcome are treated equally.

$$P(\hat{y} = 1|z = 0, y = 1) = P(\hat{y} = 1|z = 1, y = 1), \quad (3)$$

In the US employment law domain, a commonly applied discrimination measure is the 4/5 rule, a heuristic rule that states that if the hiring rate of a demographic group is less than 0.8 of the group with the highest hiring rate, there is evidence of disparate impact (Commission et al. 1979). Consequently, the rule is a direct measure of statistical parity, as it measures whether a statistical parity ratio (i.e., the ratio obtained when comparing the proportions of two demographic groups) exceeds the 0.8 threshold. This rule operates as a “rule of thumb”, since failing to achieve it is not enough to demonstrate discrimination or unfairness (Watkins, McKenna, and Chen 2022), but rather to denote a *prima facie evidence* that discrimination may be occurring. However, we can use the demographic ratio as a fairness measure in the scoring tools trained with FairCV data, as the dataset groundtruth (i.e., the blind scores y) is completely balanced in terms of gender, with equal representation of each group. This should prevent our assessment from failing to consider contextual variables related to gender that affect the proportions of women in the predictions

beyond the bias in y^G , thus preventing a portability trap in the application of this measure. Furthermore, we will use the recall computed for different groups to include a fairness measure that also considers the utility of the system.

6 Experimental Setup and Results

We train the scoring tool using the architecture of Fig. 1 using FairCVdb (Peña et al. 2023). We explore two Transformers (Sect. 5.2) for the text understanding module \mathbf{W}^T , which are frozen during training. The feature vector $\mathbf{f}^t \in \mathbb{R}^{768}$ is obtained as the average of all the output embeddings. Taking into account \mathbf{W}^F , we implement it as a naive MLP with three layers. The first two layers, with 300 and 20 units, further process \mathbf{f}^t . Then we concatenate the output of these layers with the candidate competencies \mathbf{x} , and compute the final score \hat{y} with an output layer. We used a sigmoid activation in all layers and a dropout rate of 0.3. The system is trained for 10 epochs using RMSE loss, batch size of 32, AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e - 8$), and $lr = 1e - 3$. We use the train/validation splits from FairCVtest (Peña et al. 2020a,b, 2023). All experiments were carried out on a single PC with a NVIDIA RTX 4090 (24 GB), Intel Core i9, 32GB RAM.

6.1 Gender Information Analysis

In the first experimental part, we want to assess the capabilities of the system to detect information related to the sensitive attribute z from the representations extracted by the transformer model. To do so, we propose a screening process based on 4 different resume scoring systems. The idea is to use each system to assign a score to the resumes included in the validation set, then rank them according to the predicted scores \hat{y} , and extract gender proportions in the pool of suggested candidates. Each system in this experiment was trained using one of the two Transformer models and one of the two scores available on the resumes (i.e., unbiased y and gender-biased scores y^G).

Tab. 1 presents the proportion of males and females in a shortlist with the top-500 candidates, according to the scores predicted by each system. Note that we also included the Kullback-Leibler divergence D_{KL} between the gender distributions of the predicted scores as a quantitative measure of the gender bias of the scoring tools, as previously used by Peña et al. (2023). Observing the systems trained with the unbiased scores, we can observe a similar behavior using both language models. The proportions of males and females in the shortlist of candidates are nearly equal for the RoBERTa-based model, and even perfectly equal for the BERT-based model. In this context, systems learn to predict unbiased scores, as indicated by the low values of D_{KL} across the distributions. The demographic ratio in the unbiased scenario is greater than 0.91 for both models, further highlighting the similar presence between the groups in the shortlist. In terms of recall, the BERT-based scoring tool correctly includes 78.2% of the candidates with the highest ground-truth score on its shortlist, a slightly higher value than the RoBERTa-based model (that is, 75.20%). If we observe the recall by gender, we see a gap shorter than 3 points in both cases, with a higher recall for women.

However, we observe a clear difference when training models with gender-biased scores. Both scoring systems are clearly affected by the gender bias induced in the scores. This can be seen in how D_{KL} increased over 0.3 for the distributions of the scores predicted with both models. The proportion of males is 69.4% for the BERT-based model and 67% for the RoBERTa-based model, leaving the female rates at nearly 30%. In these conditions, the demographic ratios decrease to 0.441 and 0.493, respectively. Note that, according to the 4/5 rule, a ratio lower than 0.8 is considered evidence of discrimination. Recall that gender information is not included in the candidate competencies \mathbf{x} . Therefore, the systems are extracting gender information encoded within the deep representation \mathbf{f}^t and exploiting it to fit the loss objective during training by learning the bias in the scores. In terms of recall, the higher presence of males in the shortlist led to an increase in the recall of this group, which raised about 10 points for both systems. Naturally, the opposite occurred with women, for which a drop of more than 20 points is observed, leading to a clear violation of the criterion of Equality of Opportunity. Interestingly, overall recall decreased significantly for both systems, highlighting the fact that the systems are not only discriminating against women, but recovering a lower percentage of suitable candidates.

6.2 Approach 1: Via Model Explainability

As we have shown in the previous section, when training a scoring tool when gender bias is present in the objective function, both systems establish a correlation between the scores and the gender information encoded in the text representations extracted by the Transformer. Thus, the final system shows biased behavior towards this group when predicting new scores, reproducing the bias in the training set. To mitigate the effect of such bias, here we apply an iterative learning process based on the use of the Integrated Gradients (Sundararajan, Taly, and Yan 2017) technique to detect gendered words and remove them from the input before re-training the system (see Section 4).

Starting from the biased systems that we trained in the previous section, we compute the attributions of the output with respect to the Transformers’ embedding layer, and select the tokens based on the embedding relevance. We use a zero vector as reference for the candidate competencies and a sequence of [PAD] tokens as reference for the text input in each case. For each resume, we extract the 20 most relevant tokens. We then extract the frequency of the tokens in this set of tokens by gender and labor sector, obtaining for each of the resulting 8 groups a set of the 30 most relevant tokens. To create these sets, we previously filtered stop words, adverbs, and verbs that provide little semantic information.

The first thing we notice when inspecting the sets is the presence of words related to the corresponding sector. This fact corroborates the usefulness of the method for detecting relevant words. We detect in the sets the gendered words manually removed by De-Arteaga et al. (2019). We also find several words denoting potential social biases in the sets, such as *children* in both the BERT and RoBERTa female set for the education sector, and the word *family* in the BERT

Model	Target	D_{KL}	Proportion			Recall		Overall
			M	F	Ratio	M	F	
BERT _{BASE}	Unbiased	0.0083	48.60%	51.40%	0.946	76.80%	79.60%	78.20%
	Biased	0.3102	69.40%	30.60%	0.441	87.60%	54.80%	71.20%
RoBERTa _{BASE}	Unbiased	0.0116	47.80%	52.20%	0.916	74.00%	77.20%	75.20%
	Biased	0.3387	67.00%	33.00%	0.493	83.60%	56.00%	69.80%

Table 1: Gender proportions in the top-500 candidates of the validation set of FairCVtest (Peña et al. 2020b). We obtained the shortlists of candidates by computing resume scores \hat{y} with different configurations of the scoring tool.

Model	Method	D_{KL}	Proportion			Recall		Overall
			M	F	Ratio	M	F	
BERT _{BASE}	Approach 1	0.0267	50.80%	49.20%	0.969	74.40%	76.40%	75.40%
	Approach 2	0.0422	54.20%	45.80%	0.845	86.40%	79.60%	83.00%
RoBERTa _{BASE}	Approach 1	0.0168	49.00%	51.00%	0.961	71.60%	74.40%	73.00%
	Approach 2	0.0487	54.60%	45.40%	0.831	83.40%	76.60%	80.00%

Table 2: Gender proportions in the top-500 candidates of the validation set of FairCVtest. For both scoring tools, we extract results after removing gender information in the input space (Approach 1) or in a latent space (Approach 2).

female set for the healthcare sector and in the RoBERTa female set for the jurisdiction sector. Although one could argue that *children* is a word expected in the education sector, the absence of such a word in the corresponding male set is consistent with social stereotypes that place women in a caregiving position. The same could be argued for the case of *family*. Interestingly, we also find the word *husband* in BERT’s female set for education. This word does not provide any meaning for the labor sector, and in fact can serve as a proxy for the gender of the bio (in a context where most women in the dataset are heterosexual). Other interesting words aligning with social gender stereotypes can be found in the education sector. Concretely, in the female set we encounter the words *psychology* and *health*, while in the male set we find *technology*, *engineering*, and *economics*. This distinction is in line with the social stereotypes about the prevalence of each group in different occupations. We note here that this kind of words may be those referred to in the conclusions of De-Arteaga et al. (2019), since they correlate with gender by unequal representations instead of explicitly containing gender information.

We replace the previous tokens, as well as the bios names, with the BERT/RoBERTa mask token ([MASK]), and re-train the scoring tool \mathbf{W}^S with the biased scores. In this case, both systems show an unbiased behavior in their predictions \hat{y} . Tab. 2 presents the proportion of each gender in the top-500 candidate shortlist using the scoring tool trained with this approach. Despite being trained with the same biased target function as the baseline models, both genders have almost equal representation in these lists, exhibiting even higher demographic ratios than those obtained with the unbiased models (cf. Tab. 1). By removing gendered words from the input bios, the gender information within the system was effectively reduced. Hence, trained systems cannot reproduce the bias, as they have no information about the cause (i.e., z -related membership) of the difference in

scores. We can corroborate bias mitigation by looking at the predicted score distributions \hat{y} by gender on the validation set. Although the D_{KL} between the distributions is higher than 0.3 for both baseline models, the value is reduced to 0.0267 for the BERT-based tool and 0.0168 for the RoBERTa-based one.

Looking now at the utility of the systems, both systems present a higher overall recall than the biased scoring tools. The removal of gender information not only equalized the representation of both groups, but improved the retrieval of suitable candidates. By not distinguishing the source of the bias in the target function y^G , the scoring tools are forced to pay more attention to the rest of the features in order to fit the training objective. Hence, we have improved not only the fairness of the systems, but also their utility. Note here that the recall for men decreased compared to biased scenarios. The increase in recall in men when using biased tools was due to the greater presence of this group in the candidate shortlist, rather than to better performance. This can be corroborated by comparing the gender recall values obtained here with the values obtained with the unbiased tools, which are closer. However, the recall values obtained by the unbiased tools are slightly higher, indicating that even after mitigating the effect of gender bias, this method is still penalized for it.

6.3 Approach 2: Via Adversarial Learning

In this section, we apply the method of Kim et al. (2019) (LNTL) to remove gender information in latent representations of scoring tools \mathbf{W}^S during training. Concretely, we will remove the gender information from the 300 dimensional feature vector computed by the first layer of \mathbf{W}^F . To this end, we attach an auxiliary gender classifier to this layer. The gender classifier is made up of a hidden layer with 20 units and an output layer with two output units, softmax activation. Similarly to \mathbf{W}^F , we set a dropout rate of 0.3 and

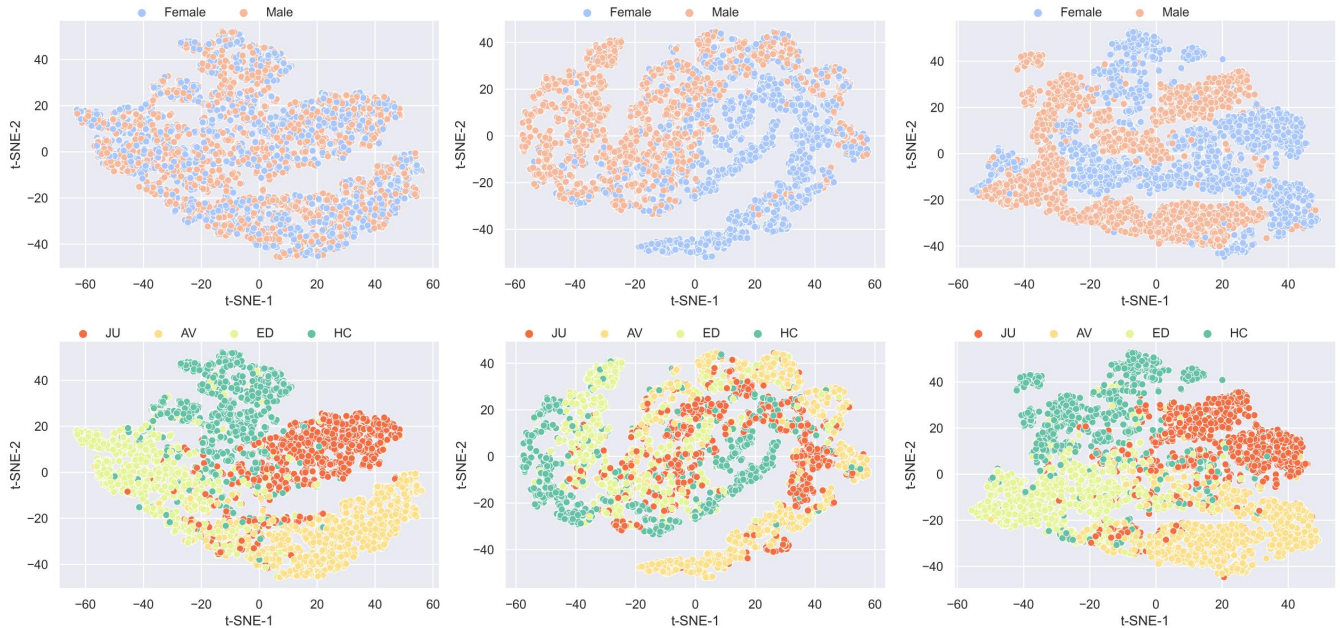


Figure 3: Visualization of the t-SNE projections for the 300-dimensional feature embeddings computed with the BERT-based scoring tool in different setups: (Left) the baseline scoring tool trained with unbiased scores; (Center) the baseline scoring tool trained with biased scores; and (Right) the scoring tool trained with biased scores using the Approach 2.

use a value of $\lambda = 0.1$ to balance the terms in Eq. 1.

The results of training the scoring tools with biased scores using this approach are presented in Tab. 2. Similarly to the results obtained with Approach 1, the recruitment systems trained with the adaptation of LNTL significantly reduced the impact of bias. Looking first at the distributions of \hat{y} , both systems exhibit a D_{KL} value under 0.05, far from those obtained in the biased baseline scenario. However, the values obtained with the approach 2 are far from the unbiased setup, and significantly higher than the ones of Approach 1. This indicates that while this method has significantly reduced the effect of bias, it has not removed it as effectively as the previous approach. However, for both systems, the demographic ratios comply with the rule of 4/5 and are significantly higher than at the baseline. Although the gender gap is greater here than when using Approach 1 or the unbiased baseline, we note that the recall values obtained are the highest among all configurations. Surprisingly, both systems achieve recall values in the male group close to those obtained with the biased baselines, without the increased presence of this group in the shortlist. This fact indicates that the removal of latent gender information encouraged the tool to focus on the rest of the features, leading to a more accurate prediction of the score.

To further analyze the impact of LNTL in latent representations, we compute the t-SNE (Van der Maaten and Hinton 2008) projections of the 300-dimensional latent vectors from the FairCV validation set. This dimensionality reduction representation works on the basis that similar samples in the original spaces should be projected to nearby points in the low-dimensional space, while dissimilar samples should

fall far apart. Thus, the representation is useful for understanding relationships and data clusters in the original space but does not have class semantic meaning, as it is an unsupervised method. We present the t-SNE projections computed for the BERT-based scoring tool in Fig. 3, for which we provide annotations of gender (Upper row) and labor sector (Bottom row) after computing the representations. Figure 3 (Center) presents the projection of the feature vectors extracted with the biased baseline, while Figure 3 (Right) presents the proper when applying LNTL. We also included in Figure 3 (left) the projections of the baseline tool trained with the unbiased scores y for comparison purposes.

Having first attended the unbiased tool, we appreciate that the representations ignore gender information, with the samples forming clusters based on the labor sector. Conversely, we observe that the labor clusters disappear in the embeddings extracted with the biased baseline tool, as the samples belonging to each labor sector appear dispersed (i.e., for instance, the Jurisdiction group), and even mixed. Instead, we can observe how most of the samples seem to group in two regions according to gender (Figure 3 (Center-Up)). Despite not being explicitly trained to infer gender, because this information is crucial to fit the biased target function, the latent representations capture gender information with even more relevance than labor-related information. When Approach 2 is applied, an interesting projection is obtained. We notice that the labor-related information regains significance within the latent representation compared to the biased scenario, as the samples seem to group again by labor sector. However, if we pay attention to the projections annotated by gender, we can appreciate that, within each labor cluster, the model

is able to distinguish each gender group. Note that, by not forming general clusters as in the biased scenario, we cannot conclude that gender is the predominant information within the latent representation, but the tool clearly has some sense of it. This is consistent with the results presented in Table 2, in light of which we have already commented that the gender information was not completely removed, as denoted by the performance gaps between groups.

7 Conclusions

We have addressed the topic of gender bias and fairness in LLMs with application to the labor market. We have used a synthetic profile dataset to evaluate the capacity of a Transformer-based scoring tool to exploit gender-related information in textual data. Our initial hypothesis was that, in the presence of a biased target function, the system would use the information in the text input to learn the bias. Hence, systems trained in this context would reproduce biased behavior toward a gender group.

We explored the behavior of two different LLMs, BERT and RoBERTa, in the context described above. In our experiments, we demonstrate how the systems based on both models effectively learn bias. We explore the application of privacy-enhancing methods to reduce gender information in the system as a way to prevent it from learning bias. Concretely, our framework adapts two different methods, a deep explainability-based technique and a bias-aware learning method. Our experiments show how the application of both methods successfully reduced gender information in the pipeline, thus preventing biased behaviors despite training in the presence of data biases.

In future work, we plan to go deeper into measuring (Serna, Morales, and Fierrez 2025), explaining (Barredo Arrieta et al. 2020; Ortega et al. 2021; Tello et al. 2023), and mitigating biases (Mancera et al. 2025b) in NLP models following related advances in Computer Vision and general AI research. We also plan to align our bias research with recent efforts trying to link already trained machine learning models with the data used for their training. A very challenging and fundamental question in many scenarios is: Has this given data been used to train this particular model? Answering this question with a certain precision has been shown to be feasible in general CV (DeAlcala et al. 2024, 2025c,a,b) and NLP setups (Mancera et al. 2025a) with state-of-the-art models. Our vision for the future is to try to advance a subsequent research question (perhaps using human-in-the-loop procedures (Peña et al. 2024)): What data used for training this particular model are the source of this observed bias?

Acknowledgments

This study has been supported by the projects BBforTAI (PID2021-127641OB-I00 MICINN/FEDER), HumanCAIC (TED2021-131787B-I00 MICINN), M2RAI (PID2024-160053OB-I00 MICIU/FEDER), and Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-2). The work G. Mancera is supported by FPI-PRE2022-104499 MICINN/FEDER. The work has been conducted within the ELLIS Unit Madrid.

References

- Acien, A.; Morales, A.; Vera-Rodríguez, R.; Bartolome, I.; and Fierrez, J. 2018. Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition. In *IAPR Iberoamerican Congress on Pattern Recognition (CIARP)*, volume 11401 of *LNCS*, 584–593. Springer.
- Altonji, J. G.; and Blank, R. 1999. Race and Gender in the Labor Market. In *Handbook of Labor Economics*, volume 3, chapter 10, 3143–3259. Elsevier.
- Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Benetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.
- Becerra, A.; Daza, R.; Cobos, R.; Morales, A.; and Fierrez, J. 2025. M2LADS Demo: A System for Generating Multimodal Learning Analytics Dashboards. In *AAAI Workshop on Innovation and Responsibility in AI-Supported Education (iRAISE)*.
- Bendick Jr, M.; Jackson, C. W.; and Romero, J. H. 1997. Employment Discrimination Against Older Workers: An Experimental Study of Hiring Practices. *Journal of Aging & Social Policy*, 8: 25–46.
- Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg more Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4): 991–1013.
- Blodgett, S.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 5454–5476.
- Bogen, M.; and Rieke, A. 2018. Help Wanted: Examination of Hiring Algorithms, Equity, and Bias. Technical report.
- Bolukbasi, T.; Chang, K.; Zou, J.; Saligrama, V.; et al. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2016)*, 4356–4364.
- Bommasani, R.; Klyman, K.; Zhang, D.; and Liang, P. 2023. Do Foundation Model Providers Comply with the EU AI Act?
- Caliskan, A.; Bryson, J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science*, 356(6334): 183–186.
- Cao, Y.; and Daumé, H. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. *Computational Linguistics*, 47(3): 615–661.
- Chen, Y.; Mahoney, C.; Grasso, I.; Wali, E.; et al. 2021. Gender Bias and Under-Representation in Natural Language Processing Across Human Languages. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 24–34.
- Cheng, L.; Varshney, K. R.; and Liu, H. 2021. Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *Journal of Artificial Intelligence Research*, 71: 1137–1181.

- Commission, U. E. E. O.; et al. 1979. Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures. *Federal Register*, 40(43).
- Dastin, J. 2018. Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women. *Reuters*.
- Daza, R.; Morales, A.; Tolosana, R.; Gomez, L. F.; Fierrez, J.; and Ortega-Garcia, J. 2023. edBB-Demo: Biometrics and Behavior Analysis for Online Educational Platforms. In *AAAI Conf. on Artificial Intelligence (AAAI)*, volume 37, 16422–1642.
- De-Arteaga, M.; Romanov, R.; Wallach, H.; Chayes, J.; et al. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 120–128.
- DeAlcala, D.; Morales, A.; Fierrez, J.; Mancera, G.; and Tolosana, R. 2025a. gMINT: Gradient-based Membership Inference Test applied to Image Models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2781–2790.
- DeAlcala, D.; Morales, A.; Fierrez, J.; Mancera, G.; Tolosana, R.; and Ortega-Garcia, J. 2025b. Active Membership Inference Test (aMINT): Enhancing Model Auditability with Multi-Task Learning. In *IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*.
- DeAlcala, D.; Morales, A.; Fierrez, J.; Mancera, G.; Tolosana, R.; and Vera-Rodriguez, R. 2025c. MINT-Demo: Membership Inference Test Demonstrator. In *AAAI Workshop on AI Governance: Alignment, Morality, and Law (AIGOV)*.
- DeAlcala, D.; Morales, A.; Mancera, G.; Fierrez, J.; Tolosana, R.; and Ortega-Garcia, J. 2024. Is my Data in your AI Model? Membership Inference Test with Application to Face Images. *arXiv preprint arXiv:2402.09225*.
- DeAlcala, D.; Serna, I.; Morales, A.; Fierrez, J.; et al. 2023. Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1167–1172.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, 4171–4186.
- Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; et al. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 67–73.
- Executive Office of the President. 2016. Big Data: A report on Algorithmic Systems, Opportunity, and Civil Rights. Technical report, The White House.
- Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.
- Ge, S.; Zhou, C.; Hou, R.; Khabsa, M.; et al. 2023. MART: Improving LLM Safety with Multi-round Automatic Red-Teaming. Preprint at <https://arxiv.org/abs/2311.07689>.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; et al. 2014. Generative Adversarial Nets. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014)*, volume 27, 2672–2680.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 3323–3331.
- Haryan, S.; Malik, R.; Redij, P.; et al. 2024. FairHire: A Fair and Automated Candidate Screening System. In *International Conference on Machine Intelligence, Tools, and Applications*, 372–382. Sehore, India: Springer Nature Switzerland.
- Hassanpour, A.; Moradikia, M.; Yang, B.; Abdelhadi, A.; et al. 2022. Differential Privacy Preservation in Robust Continual Learning. *IEEE Access*, 10: 24273–2428.
- Hu, E.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceeding of the International Conference on Learning Representations (ICLR)*.
- Keyes, O. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of ACM Human-Computer Interaction*, 2(CSCW): 1–22.
- Kim, B.; Kim, H.; Kim, K.; Kim, S.; et al. 2019. Learning Not to Learn: Training Deep Neural Networks with Biased Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 9012–9020.
- Kiritchenko, S.; and Mohammad, S. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 43–53. New Orleans, Louisiana: Association for Computational Linguistics.
- Leavy, S.; Siapera, E.; and O’Sullivan, B. 2021. Ethical Data Curation for AI: An Approach Based on Feminist Epistemology and Critical Theories of Race. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 695–703.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv/1907.11692*.
- Mancera, G.; DeAlcala, D.; Fierrez, J.; Tolosana, R.; and Morales, A. 2025a. Is My Text in Your AI Model? Gradient-based Membership Inference Test applied to LLMs. In *arXiv preprint arXiv:2503.07384*.
- Mancera, G.; Morales, A.; Fierrez, J.; et al. 2025b. PBA-LLM: Privacy-and Bias-aware NLP using Named-Entity Recognition (NER). In *Proceedings of IAPR International Conference on Document Analysis and Recognition Workshops (ICDARw)*.
- May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the Conference of the*

- North American Chapter of the Association for Computational Linguistics: *Human Language Technologies (NAACL-HLT)*, 622–628.
- Melzi, P.; Tolosana, R.; Vera-Rodriguez, R.; Kim, M.; Rathgeb, C.; Liu, X.; DeAndres-Tame, I.; Morales, A.; Fierrez, J.; Ortega-Garcia, J.; et al. 2024. FRCSyn-onGoing: Benchmarking and Comprehensive Evaluation of Real and Synthetic Data to Improve Face Recognition Systems. *Information Fusion*, 107: 102322.
- Morales, A.; Fierrez, J.; Vera-Rodriguez, R.; and Tolosana, R. 2021. SensitiveNets: Learning Agnostic Representations with Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 2158–2164.
- OpenAI. 2023. GPT-4 Technical Report. *arXiv/2303.08774*.
- Ortega, A.; Fierrez, J.; Morales, A.; Wang, Z.; Cruz, M.; Alonso, C. L.; and Ribeiro, T. 2021. Symbolic AI for XAI: Evaluating LFIT Inductive Programming for Explaining Biases in Machine Learning. *Computers*, 10(11): 154.
- Park, J.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2799–2804.
- Peña, A.; Morales, A.; Fierrez, J.; Serna, I.; et al. 2023. Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs. In *Document Analysis and Recognition – ICDAR 2023 Workshops*, 20 – 33.
- Peña, A.; Fierrez, J.; Lapedriza, A.; and Morales, A. 2021. Learning Emotional-Blinded Face Representations. In *IAPR Intl. Conf. on Pattern Recognition (ICPR)*, 3566–3573.
- Peña, A.; Morales, A.; Fierrez, J.; Ortega-Garcia, J.; Puente, I.; Cordova, J.; and Cordova, G. 2024. Continuous document layout analysis: Human-in-the-loop AI-based data curation, database, and evaluation in the domain of public affairs. *Information Fusion*, 108: 102398.
- Peña, A.; Serna, I.; Morales, A.; and Fierrez, J. 2020a. Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRw)*. Also presented at ICML 2020 Workshop on Human-in-the-Loop Learning.
- Peña, A.; Serna, I.; Morales, A.; and Fierrez, J. 2020b. Fair-CVtest Demo: Understanding Bias in Multimodal Learning with a Testbed in Fair Automatic Recruitment. In *ACM Intl. Conf. on Multimodal Interaction (ICMI)*, 760–761. Also presented at European Conf. on Computer Vision, ECCV 2020.
- Peña, A.; Serna, I.; Morales, A.; Fierrez, J.; et al. 2023. Human-Centric Multimodal Machine Learning: Recent Advances and Testbed on AI-based Recruitment. *SN Computer Science*, 4(43).
- Prates, M.; Avelar, P.; and Lamb, L. 2020. Assessing Gender Bias in Machine Translation: A Case Study with Google Translate. *Neural Computing and Applications*, 32: 6363–6381.
- Qian, R.; Ross, C.; Fernandes, J.; Smith, E. M.; et al. 2022. Perturbation Augmentation for Fairer NLP. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9496–9521.
- Rahwan, I.; Cebrian, M.; Obradovich, N.; et al. 2019. Machine Behaviour. *Nature*, 568: 477–486.
- Romanov, A.; De-Arteaga, M.; Wallach, H.; Chayes, J.; et al. 2019. What’s in a Name? Reducing Bias in Bios without Access to Protected Attributes. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, 4187–4195.
- Rudinger, R.; May, C.; and Van Durme, B. 2017. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 74–79.
- Sánchez, J.; Dencik, L.; and Edwards, L. 2020. What Does It Mean to ‘Solve’ the Problem of Discrimination in Hiring? Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 458–468.
- Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; et al. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1668–1678.
- Schumann, C.; Foster, J. S.; Mattei, N.; and Dickerson, J. P. 2020. We Need Fairness and Explainability in Algorithmic Hiring. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 1716–1720.
- Serna, I.; DeAlcala, D.; Morales, A.; Fierrez, J.; and Ortega-Garcia, J. 2022a. IFBiD: Inference-Free Bias Detection. In *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, volume 3087 of *CEUR*.
- Serna, I.; Morales, A.; and Fierrez, J. 2025. Unraveling Machine Behavior by Multi-Level Bias Analysis and Detection: Methodology and Application to Face Biometrics. *arXiv preprint*.
- Serna, I.; Morales, A.; Fierrez, J.; Cebrian, M.; Obradovich, N.; and Rahwan, I. 2020. Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics. In *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, volume 2560 of *CEUR-WS*, 146–152.
- Serna, I.; Morales, A.; Fierrez, J.; and Obradovich, N. 2022b. Sensitive Loss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning. *Artificial Intelligence*, 305: 103682.
- Serna, I.; Peña, A.; Morales, A.; and Fierrez, J. 2021. Inside-Bias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics. In *Proceedings of IAPR International Conference on Pattern Recognition (ICPR)*, 3720–3727. Milan, Italy.
- Sheng, E.; Chang, K.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3407–3412.

Singla, P.; Kaur, J.; Anju, et al. 2024. Streamlining Talent Acquisition: A Machine Learning Approach to Automated Resume Screening. In *2024 Second International Conference on Advanced Computing Communication Technologies (ICACCTech)*, 69–75. Sonipat, India: IEEE.

Stanovsky, G.; Smith, N.; and Zettlemoyer, L. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1679–1684.

Stragapede, G.; Vera-Rodriguez, R.; Tolosana, R.; Morales, A.; Damer, N.; Fierrez, J.; and Ortega-Garcia, J. 2024. Keystroke Verification Challenge (KVC): Biometric and Fairness Benchmark Evaluation. *IEEE Access*, 12: 1102–1116.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 3319–3328. PMLR.

Tello, J.; de la Cruz, M.; Ribeiro, T.; Fierrez, J.; Morales, A.; Tolosana, R.; Alonso, C. L.; and Ortega, A. 2023. Symbolic AI (LFIT) for XAI to Handle Biases. In *European Conf. on Artificial Intelligence Workshops (ECAIW)*, volume 3523 of CEUR-WS.

Terhorst, P.; Kolf, J. N.; Huber, M.; Kirchbuchner, F.; Damer, N.; Morales, A.; Fierrez, J.; and Kuijper, A. 2022. A Comprehensive Study on Face Recognition Biases Beyond Demographics. *IEEE Trans. on Technology and Society*, 3(1): 16–30.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; et al. 2023. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. Technical report, Meta AI.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; et al. 2017. Attention is All You Need. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, volume 30, 6000–6010.

Watkins, E. A.; McKenna, M.; and Chen, J. 2022. The Four-Fifths Rule is not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness. *arXiv/2202.09519*.

Yang, K.; et al. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the International Conference on Scientific and Statistical Database Management*, 1–6.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. FA*IR: A Fair Top-K Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1569–1578.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 2, 15–20.