Analyzing bias and discrimination in an algorithmic hiring use case

David Quer, Anna Via, Marc Serra-Vidal, Laia Nadal and Didac Fortuny

Abstract

Algorithmic hiring, powered by AI, has become prevalent in recruitment processes. This paper investigates bias and discrimination in a specific Machine Learning use case at InfoJobs, the leading recruitment platform in Spain. The motivation stems from ethical, legal, and reputational considerations, emphasizing the importance of building responsible and fair AI systems in recruitment. The study presents a comprehensive analysis, employing fairness metrics and, additionally, a novel Granular Bias Measure (GBM) introduced to assess biases at the individual job title level. The experiments involve real candidate experiences, and the results indicate that there is no significant negative discrimination towards female experiences. However, opportunities for improving model results for specific age groups, such as <25, are identified. The paper concludes by highlighting the need for further analysis to understand and address biases effectively. While bias detection is achievable, addressing root causes and ensuring fairness requires ongoing monitoring and improvements in algorithmic hiring systems.

1. Motivation

1.1. Company context

Algorithmic hiring refers to the use of algorithms, often powered by artificial intelligence (AI) and machine learning (ML) technologies, in the recruitment and hiring processes. The end goal of leveraging these technologies is to optimize and simplify processes within the Job Boards platforms. According to [1], around 99% of Fortune 500 companies use talent-sifting software in some part of the recruitment and hiring process. Due to the usage of ML models and potential automation of decisions, the risks related to bias and discrimination in Algorithmic Hiring are huge ([2], [3]).

InfoJobs[1] is the leading recruitment platform in Spain and part of Adevinta group[2] . The platform dealt with over 2.7M offers and more than 125M applications in 2022. This big volume and scale implies it is needed to find ways to optimize efficiency and value for candidates and recruiters. For that reason, the platform incorporates Machine Learning systems in several touchpoints of the user and recruiter experience such as offer recommendations for candidates (to help candidates find relevant job offers), automatic CV parsing from PDF (to help candidates update

AIMMES 2024 Workshop on AI bias: Measurements, Mitigation, Explanation Strategies | co-located with EU Fairness Cluster Conference 2024, Amsterdam, Netherlands

[1] https://www.infojobs.net/

[2] https://adevinta.com/

the InfoJobs profile with information available in a PDF), and offer-CV matching computation (to help candidates and recruiters assess fitness between a CV and an offer).

As these ML systems can influence users' and recruiters' decisions around employment, many risks arise with algorithmic hiring. It makes it mandatory to find ways to ensure these systems are built responsibly, avoiding biases and discrimination, and bringing a positive impact to society:

- **Ethically**, it is essential to prevent the perpetuation of discrimination and injustice in recruitment.
- **From a legal standpoint**, adherence to fairness principles is crucial to avoid legal consequences and comply with regulations (especially with the upcoming AI Act, where recruitment is considered High Risk).
- **From a company reputation point of view**, creating fair and transparent AI systems builds user trust and contributes to a positive corporate reputation.

1.2. Project context

As explained in [4], Natural Language Processing (NLP) techniques are widely used in algorithmic hiring to standardize candidate experiences and job offers, and align on mutual relevancy. In this paper, we present a bias and discrimination analysis for a specific use case in InfoJobs, Job Title normalizations. Job Title normalizations are core in the platform, as they allow a better understanding and standardization of offers and experiences, and nurture many functionalities like offer recommendations to candidates, offers search, alerts, or offer-cv matching.

This normalization model is implemented using the BERT language model [5], which, once fine-tuned with InfoJobs data, allows the classification of offers and candidate experience into a common Job Title taxonomy, based on ESCO [3] occupations taxonomy. This taxonomy provides descriptions of 3008 occupations and 13.890 skills linked to these occupations, translated into 28 languages. The main modifications to the ESCO taxonomy consisted on:

- Ensuring all gender forms for a job title appear (e.g. "camarero", "camarera", "camarero/a", "camarero/camarera"), and the primary form is inclusive ("camarero/a").
- Gathering feedback from users and manually revising the taxonomy to improve it.

Due to the importance of this model within the platform, an analysis was prioritized to measure, detect, and mitigate potential sources of bias in the normalization of experiences. As due to GDPR the only sensitive attributes a company can store are gender and age, analysis where limited to this two attributes and their intersectionality.

2. Studies, State-of-the-art

Fairness measures can target different properties. According to *Fabris et al.* survey on Fairness and Bias in Algorithmic Hiring [6], fairness is a broad concept which can be divided in Procedural

[3] https://esco.ec.europa.eu/en/classification

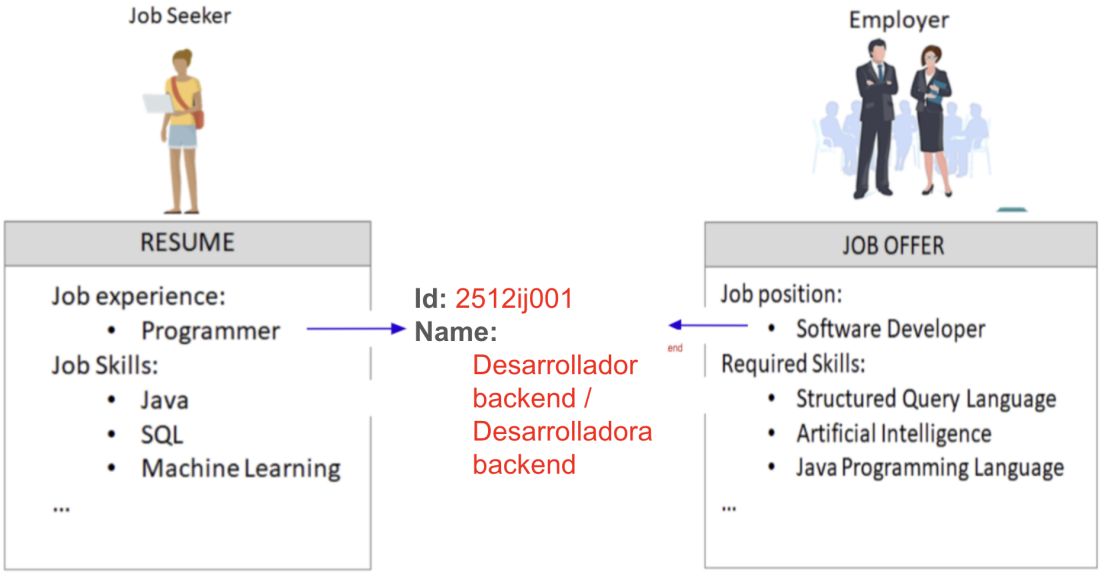

Figure 1: Example of experience "Programmer" and offer "Software Developer" normalizating into the same "Back End Developer" Job Title

fairness, Outcome fairness, Accuracy fairness, Impact fairness and Representational fairness.

Impact fairness and Process fairness are related to the screening and selection processes, which are out of the scope of our work. Also Representational fairness, which focuses on the wording of the offers is not dealt with in this paper.

In this work, in order to measure biases for our model, the Job Title classification algorithm, we will focus on Outcome fairness, that looks at predictions from the perspective of candidates and Accuracy fairness, that requests equalization of accuracy-related properties between groups from the decision making perspective.

The survey [6] describes several fairness metrics for evaluating algorithmic systems, primarily focused on ranking and classification tasks. When considering our AI system, a classification problem narrows down the metrics that we can use. As for Outcome fairness metrics the survey [6] describes as suitable metrics to evaluate fairness: Disparate Impact (DI), Demographic Disparity (DD), Representation in Positive Predicted (RPP) and True Positive Rate Difference (TPRD).

Having said that, DI, requires a notion of the best and worst-off groups to be interpretable and only considers two groups, does not account for non-binary attributes and focuses just on ratio, not absolute rates, DD is prone to masking effect when acceptance rates are very small, making them unsuitable for this analysis. As for RPP, it is used in cases where the overall population of interest for an algorithm is unknown and that it's not our case.

Because of these limitations in the other metrics, we chose the True Positive Rate Difference

(TPRD) measure, which measures disparities in true positive rates (also known as recall) between different sensitive groups and closely related to equal opportunity and separation.

To have a better understanding of the Biases that our algorithm might have, we have considered also explaining away with conditional probabilities, True Positive Rate and True Negative Rate, before calculating the True Positive Rate Difference (TPRD) measure.

With respect to Accuracy fairness metrics we picked up Balanced Classification Rate Difference (BCRD), that is a measure of the disparity in classification accuracy between groups because the other metric proposed in the survey [6], as Mean Absolute Error (MAE) targets only regression problems.

Finally, given that the Job Title classification algorithm is a multiclass classification model we believed a deeper understanding was needed for every Job Title prediction, and we built the Granular Bias Measure (GBM), based on Prevalence [7], to do so.

3. Experiments

As the goal was to assess the real, production fairness of the normalization predictions, we used real candidate experiences and their Job Title normalization predictions. We focused our efforts on the predictions (output) leaving aside a proper analysis of the unbalance of the training data (input).

3.1. Data input

Nevertheless, we have done some sanity checks in terms of proportions to be sure that we don't have skewed data.

Considering Male vs Female genders, we have 55 vs 45% of data, and when intersecting gender with the age groups of <25, 25-45 and >45 years old, no group represents less than 5%, being Male and <25 the less represented group (9%).

3.2. Score analysis

For the Job Title Normalization outputs a score assigned for each class that represents the model's confidence in predicting a class and softmax layers in the neural network generate these probabilities. The distribution of these scores across different instances in the dataset can provide insights into the model's behavior and reliability.

As a sanity check, before analyzing the classifier, we follow the basic process of obtaining the predicted scores, construct the cumulative distribution functions for the predicted scores for each sensible variable we want to compare: Gender and Age. Finally, we apply a Kolmogorov-Smirnov test [8] to statistically assess whether the predicted scores from different classes of the sensible variable follow the same underlying distribution.

Table 1
Kolmogorov-Smirnov test results for sensitive variables

Sensitive variable		Kolmogorov-Smirnov test		
		Statistic	p-value	Accept H0 (=)
Gender	**Male vs Female**	0.41	6.16E-05	FALSE
	>45 vs 25-45	0.02	1.06E-01	TRUE
Age	**>45 vs <25**	0.11	5.45E-08	FALSE
	25-45 vs <25	0.09	5.79E-06	FALSE

Indeed, we conclude from these results that there are differences in the distributions between classes. Although Kolmogorov-Smirnov is a valuable tool for comparing the performance and calibration of models in a statistical manner, it does not tell us anything about the magnitude of the differences and which classes are more affected.

On top of that, when dealing with a classifier model the threshold selected to decide the class can render these differences insignificant. That's why we moved to proper classification metrics in order to further evaluate the fairness of our algorithm.

3.3. Classifier metrics

The Job Title Normalization model is a multiclass classification model that works as follows: the class with the highest predicted probability is obtained and, from it, like a binary classification model, applies a threshold to the predicted scores to consider a prediction valid or invalid. Thanks to a labeling technique, detailed below, we are able to assess the correctness of the predictions, thus obtaining metrics such as "estimated accuracy" across all classes.

To be able to evaluate the model as a binary classifier and calculate the metrics True Positive Rate Difference (TPRD), Balanced Classification Rate Difference (BCRD) and the explaining away with conditional probabilities, we require ground truth data and knowledge of ground truth labels.

The labeled ground truth for validation is obtained by filtering the experiences where the titles are exactly the same as one of our modified ESCO taxonomy job title groups or their synonyms. This poses the issue that, when evaluating the model, we can only do it for the datapoints for which this ESCO ID is known, but when we weight the cost of obtaining manually labeled data versus using ESCO to validate our predictions, the latter option makes it much easier for us to evaluate the model. Still, we are able to label with this technique over 50% of the data points.

3.4. Explaining away with conditional probabilities

Explaining away in the context of a confusion table involves understanding how the occurrence or absence of one event can influence the probability of another event. In a confusion table, which is commonly used to assess the performance of a classification model, explaining away refers to the impact of one class prediction on the probability of another.

When considering our sensitive variables we built these confusion tables to do the explaining away based on the criteria, what is the probability of getting a right or wrong prediction given we belong to a class (e.g. Male or Female) allowing us to see differences between classes and also guessing if the model is doing better when it gets it right.

For the variable Gender: Male or Female we get the following results:

Table 2
Confusion matrices for the variable Gender

Male	predicted	
actual	valid	invalid
correct	12230	3
incorrect	61	22

Female	predicted	
actual	valid	invalid
correct	14422	5
incorrect	249	32

In terms of True Positive and True Negative Rates by Gender group we have:

Table 3
True Positive Rate and True Negative Rate by Gender

Gender	TPR	TNR
Male	0.99975	0.26506
Female	0.99965	0.11387

For the variable Age: 25-45, <25, >45 we get the following results:

Table 4
Confusion matrices for the variable Age

25-45	predicted	
actual	valid	invalid
correct	15612	5
incorrect	129	25

<25	predicted	
actual	valid	invalid
correct	4368	1
incorrect	21	2

>45	predicted	
actual	valid	invalid
correct	6672	3
incorrect	160	27

In terms of True Positive and True Negative Rates by Age group we have:

Table 5
True Positive Rate and True Negative Rate by Age

Age	TPR	TNR
25-45	0.99967	0.16233
<25	0.99977	0.08695
>45	0.99955	0.14438

As for TPRD and BCRD, strictly speaking, this metrics are used to compare two groups which, in the case of Gender works fine but in the case of Age implies having to do the pair comparisons

for each group so, we did the Explaining Away interpretation for both sensitive variables and we calculate TPRD and BCRD only for Gender.

3.5. True Positive Rate Difference (TPRD)

For Outcome fairness we use True Positive Rate Difference (TPRD) as a metric to summarize the prediction capacity of our model and evaluate possible biases. By definitions this metric is calculated using the True Positive Rate for each group as follows:

$$TPR_g = Pr(\bar{y} = 1|y = 1, s = g) \tag{1}$$

$$TPRD = TPR_g - TPR_g c \tag{2}$$

Using the previously calculated TPR by Gender we get that the True Positive Rate Difference between Gender groups is 0.0001.

$$TPRD = TPR_{male} - TPR_{female} = 0.99975 - 0.99965 = 0.0001 \tag{3}$$

3.6. Balanced Classification Rate Difference (BCRD)

For Accuracy fairness we use Balanced Classification Rate Difference (BCRD) that checks for disparate accuracy between groups as is defined as follows:

$$BCR_g = \frac{TPR_g + TNR_g}{2} \tag{4}$$

$$BCRD = BCR_g - BCR_g c \tag{5}$$

Using the previously calculated TPR and TNR by Gender we calculate the Balanced Classification Rate by Gender group.

$$BCR_{male} = \frac{0.99975 + 0.26506}{2} = 0.63240 \tag{6}$$

$$BCR_{female} = \frac{0.99965 + 0.11387}{2} = 0.55676 \tag{7}$$

And doing the difference between Gender groups we get a Balanced Classification Rate Difference of 0.07563

$$BCRD = BCR_{male} - BCR_{female} = 0.63240 - 0.5917 = 0.07563 \tag{8}$$

And given this difference in accuracy by Gender group we decide to take a closer look by introducing the Granular Bias Measure (GBM).

3.7. Calculating a Bias Measure for every Job Title classified

Up until this point we've been using well known statistical techniques and algorithmic hiring metrics for the Bias calculation. These measures are aggregated measures representing outcome and accuracy biases but we are interested in understanding if a potential source for this bias could be the unequal gender representation in each Job Title (and the fact some will be easier to predict than others).To that end we need to look at every Job Title separately and check if the bias effect is present for each particular one. For this reason we embarked on building the Granular Bias Measure (GBM).

To build the Granular Bias Measure (GBM), that summarizes how well is classified a Job Title, we start by calculating measure inspired on Prevalence, that is the proportion of a particular population found to be affected by an outcome, for each level of the sensitive variable.

That gives us the P+, positive prevalence, when the Job Title is well classified. Note that both measure P+ is expressed as a percentage of the total population. (TP - True Positives)

$$P+_{male} = \frac{TP_{male}}{Total_{male}} \tag{9}$$

$$P+_{female} = \frac{TP_{female}}{Total_{female}} \tag{10}$$

Then we calculate the GBM measure by subtracting the prevalence from each group:

$$GBM = P+_{male} - P+_{female} \tag{11}$$

And finally to express the Bias we evaluate if the absolute value of the difference abs(GBM+) is greater than 0.0, pointing us to a possible bias. Let's see some examples:

Table 6
Sample results of GBM measure calculation on Job Titles normalized

Job Title Normalized	Total (M)	Total (F)	P+ (M)	P+ (F)	GBM
Abogado/a	54	45	1.00000	1.00000	0.00000
Agente comercial	57	51	0.96491	0.88235	0.08256
Jefe/a de compras	53	54	0.94340	0.94444	-0.00105
Administrativo/a	118	516	0.822034	0.945736	-0.123703
Cajero/Cajera	61	424	0.934426	0.985849	-0.051423
Director/a de recursos	2	1	0.50000	1.00000	-0.50000

As a result we are able to calculate the GBM for each Job Title normalized and by considering larger differences in the score extract more detailed conclusions.

4. Discussion

Model scores KS test results in table 1 show that the score distributions between gender and age groups are mostly different (except for the comparison between age groups 25-45 and >45). Also, when intersecting gender and age, we can see how these differences are maintained but don't grow through intersectionality.

We continue digging for more insightful differences with explaining away conditional probabilities which helps unveil the interdependence between different prediction outcomes.

Considering Gender sensitive variable outcomes in table 3 we can see the probability of being Male or Female does not change much between groups when the prediction is "correct" (True Positive). On the other hand, when evaluating "incorrect" predictions (True Negative) it seems that there's a higher probability of being Male if the model gets it right discarding a wrong Job Title Normalized although the volume is very small affecting very few candidates.

In the same way, when considering Age sensitive variable outcomes in table 4, although the probability of being Male or Female doesn't change much between groups it seems that <25 is little bit different from the other two 25-45 and >45 in terms of TPR and TNR, which aligns with previous conclusions from table 1.

Now we are able to calculate the previously selected metrics. As for Outcome Fairness TPRD we can see in equation (3) that the difference between Gender groups is pretty small, almost non significant. And for Accuracy Fairness BCRD, where we try to measure differences in accuracy between Gender groups and, although it's hard to interpret, doesn't seem to be affecting the model predictions.

Finally, to validate our proposed metric Granular Bias Difference, equation (8), based on the positive Prevalence of each group, equations (6) and (7), some interesting results detected are:

Almost all model predictions do not present any biases, and the GMB metric is distributed in the interval [-0.25, 0.25] showing small differences on Gender classes. Closely examining the ones that are biased can be explained for various sociological reasons.

When the number of observations by Gender class is well balanced GBM metric captures the differences between classes (see. "Abogado/a", "Agente Comercial" or "Jefe de compras" on table 6)

GBM also works well when there's an imbalance, up to a certain point, but we have enough volume of observations (see "Administrativo/Administrativa" or "Cajero/Cajera" on table 6)

For many classes, there are not enough observations to get a significant GBM, and no conclusions can be extracted from there (see. "Director de recursos" on table 6)

More work has to be done in that sense to have a more broader metric.

5. Conclusions

To conclude, we don't observe discrimination by gender in our analysis. We believe this is in part thanks to the work related to ensuring all gender forms appear in the taxonomy and have the same weight (contrary to the default male naming typical in Spanish language).

Most of the candidates seem to get a correct prediction regardless of their Gender. While it is possible a pretty small bias towards classifying properly male Job Titles only affects very few candidates and do not seem a thread on the overall fairness.

As for Age, it seems there is an opportunity to improve the model results for the <25 group that seem to be behaving differently from the other groups 25-45 and >45. We believe an explanation for that can be the fact that experiences from young people can differ from the typical experiences in ESCO taxonomy (internships, junior, leisure, private tutor...). .

Nothing indicates intersactionality between gender and age amplifies observed differences in any of the analysis.

Based on the TPRD metric we can conclude that there's no significant difference between Gender groups and there are no disparities in true positive rates. Being this measure is closely related to equal opportunity and separation we can also say that our model is giving equal opportunities independent of the Gender.

In terms of accuracy, we can conclude that, based on the metric, there is no significant difference between Gender groups. Although a closer examination will be useful to detect particular cases and infer the root causes.

Detecting or monitoring a Bias (a.k.a correlation) does not seem the harder task, but assessing the root cause of this Bias (a.k.a. causality) and correcting it is much harder and requires further analysis.

When considering the new Granular Bias Measure, GBM, more work has to be done to build a consistent and reliable summarization of the Bias. When the classes are unbalanced the GBM is very reliant on the population sample used to evaluate. Works for multiclass classification and does it better with large samples. Playing with the threshold when classifying if a GMB score is a true bias or not could yield better results.

References

[1] I. I. for the future work, Algorithmic hiring systems: what are they and what are the risks?, 2022. URL: https://www.ifow.org/news-articles/algorithmic-hiring-systems.

[2] Oleeo, Inclusive diversity in hiring guide, 2023. URL: https://www.oleeo.com/inclusive-diversity-in-hiring-guide.

[3] P. M. Kline, E. K. Rose, C. R. Walters, Systemic discrimination among large u.s. employers, 2022. URL: https://www.nber.org/system/files/working_papers/w29053/w29053.pdf.

[4] H. Kavas, M. Serra-Vidal, L. Wanner, Job offer and applicant cv classification using rich information from a labour market taxonomy, 2023. URL: http://dx.doi.org/10.2139/ssrn.4519766.
[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805.
[6] A. Fabris, N. Baranowska, M. J. Dennis, P. Hacker, J. Saldivar, F. Z. Borgesius, A. J. Biega, Fairness and bias in algorithmic hiring, 2023. URL: https://arxiv.org/abs/2309.13933.
[7] N. P. Jewell, Statistics for Epidemiology, Chapman Hall, London, 2003.
[8] Y. Z. Vance W. Berger, Kolmogorov–Smirnov Test: Overview, John Wiley Sons, Ltd., Hoboken, New Jersey, 2014.