ALGORITHMIC BIAS, FAIRNESS AND ETHICS

Mulham Fawakhrji

December 18, 2019

Abstract

This report for ALGORITHMIC BIAS, FAIRNESS AND ETHICS course assignment, after studying the Master thesis (Evaluating potential biases in commercial people search engines) and checking some of the results in the thesis, we study fairness in ranking with respect to different attributes(number of stars, and number of reviews), then we study the effect off missing values on fairness algorithms and we tried to handle it in different ways, we also try to build an initial view about the notion of more protected by combining two attributes in the dataset, finally we study the exposure concepts based on the position bias, all experiaments performed based on the two datasets (job candidates from Linkedin and top Doctors in medical specialists)[1] .

1 Fairness in ranking with respect to different attributes

In all experiments we used $FA*IR$ algorithm [2], for producing the Fair Top-k Ranking, where k subset candidates from a large pool of $n >> k$ candidates. The FA*IR algorithm receives as input a Boolean list where each position indicates if the person is part of the protected group or not, a true value in that list means that the person belongs to a protected group and false that it does not. For this analysis, every person in the dataset that has a null value in the field used to determine whether or not it is part of a protected group, has been discarded. The gaps in the dataset have been maintained so as to respect the ranking order. Afterwards, the value of k is calculated by getting the minimum amount of people left per query, data source and country bearing in mind that this value should be always greater than 10. The queries whose number of people has been lower than 10 have also been ignored. finally, different values of p have been chosen from 0.1 to 0.8 in intervals of 0.1 along with the extreme values 0.02 and 0.98. The output of the algorithm for the different values of p, k and fixed will decide if the ranking is fair enough with the protected group, query, data

Query	p = True					
	Colombia		Spain		Mexico	
	$numStar > 1$	$numStar < 1$	$numStar > 1$	$numStar < 1$	$numStar > 1$	$numStar < 1$
cirugia ora	0.3	0.8	0.7	0.5	0.6	0.6
dermatologia	0.8	0.3	0.98	0.1	0.9	0.2
ginecologia	0.8	0.4	0.9	0.2	0.9	0.3
neurocirugia	0.5	0.4	0.9	0.2	0.8	0.4
oftalmologia	0.8	0.4	0.9	0.3	0.9	0.3
otorrinolar	0.7	0.5	0.9	0.3	0.98	0.1
pediatria	0.6	0.7	0.9	0.2	0.5	0.7
psicologia	0.6	0.6	0.7	0.6	0.7	0.5
psiquiatria	0.5	0.6	0.8	0.4	0.7	0.4
urologia	0.9	0.1	0.9	0.2	0.8	0.3
Avarage	0.66	0.48	0.86	0.30	0.78	0.42

Table 1: Maximum value of p retrieved by the FA*IR algorithm considering as protected grouped the Doctor with number of stars, more and less than one, k=20 for all countries.

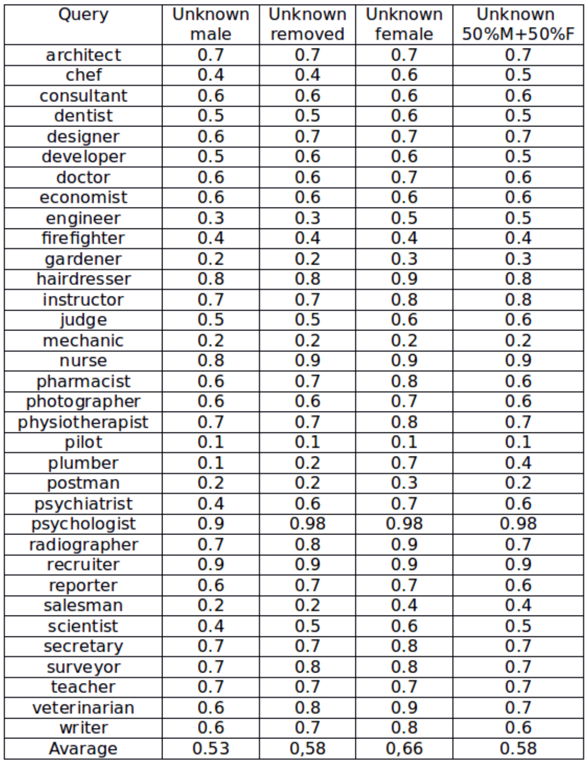
source and country given. The results the highest value of p whose value is true, meaning that it is fair, will be shown. We have studied the following protected groups, all doctors with number of star less and more than one, doctors with number of reviews less and more than one.
Table 1 shows maximum value of p retrieved by the FA*IR algorithm considering as protected grouped the Doctor with number of stars more and less than one, first we can notice in all countries that the average of p when the number of star is more than one is greater than average of p when number of star less than one with a big difference and this show majority of medical specialist is favouring doctors with number of star more than 1, except in some medical speciality there is no effect for the number of star like *cirugiaora*, and *psicologia* and in other cases has a big effect like *urologia*.
Table 2 shows maximum value of p retrieved by the FA*IR algorithm considering as protected grouped the Doctor with number of reviews more and less than one, first we can notice in Colombia in some speciality there is big difference in positive direction (number of reviews more than 1) like (urologia) and in other there is big difference in nagative direction like(cirugia ora)but in general there is no difference in average of p when the number reviews more and less than one (almost 0.01). On the other side we can see in Spain the big difference in the average of P and this show the effect of the number of reviews in the profile.

Query	p = True					
	Colombia		Spain		Mexico	
	$numRev > 1$	$numRev < 1$	$numRev > 1$	$numRev < 1$	$numRev > 1$	$numRev < 1$
cirugia ora	0.2	0.8	0.7	0.5	0.4	0.6
dermatologia	0.5	0.3	0.9	0.1	0.8	0.2
ginecologia	0.5	0.4	0.9	0.2	0.6	0.3
neurocirugia	0.4	0.4	0.9	0.2	0.6	0.4
oftalmologia	0.7	0.4	0.8	0.3	0.6	0.3
otorrinolar	0.5	0.5	0.9	0.3	0.9	0.1
pediatria	0.3	0.7	0.9	0.2	0.3	0.7
psicologia	0.3	0.6	0.7	0.6	0.5	0.5
psiquiatria	0.5	0.6	0.8	0.4	0.7	0.4
urologia	0.8	0.1	0.8	0.2	0.7	0.3
Avarage	0.47	0.48	0.83	0.30	0.61	0.42

Table 2: Maximum value of p retrieved by the FA*IR algorithm considering as protected grouped the person with number of reviews, more and less than 1, k=20 for all countries.

2 Effect of missing values on fairness

When fairness is taken into consideration, one must realise that the missing data might not be evenly distributed between different groups, which in turn might lead unwanted effects on the fairness of the data and models created, depending on how the missing data are handled, more missing values on a sensitive group than the other. People might intentionally omit information as a natural cooping mechanism when there is a belief that a truthful and complete answer might lead to a discriminatory and unfair decision. In this test we shows maximum value of p retrieved by the FA*IR algorithm considering the protected group the female gender. We tried to check the effect of missing value in gender field in the ranking on FA*IR algorithm. We consider 4 situations shown in table in figure [1], the first one every person in the dataset that has a null value in the gender field, has been discarded. The gaps in the dataset have been maintained so as to respect the ranking order we mentioned to this case in the table as (Unknown removed), in the second one we consider all unknown values in gender field as male, we mentioned to this case in the table as (Unknown male), in the third case we consider all unknown values in gender field as female, we mentioned to this case in the table as (Unknown female), in the last one we count the number of unknowns value in each profession and we consider 50% of unknown value as male and 50% as female, we mentioned to this case in the table as (Unknown 50%M+50%F). We perform our study on linked in UK. In general we can see in figure [1] that the best results in average of p was when consider all unknown field as female (0.66), and the worst case when we consider unknown field as female (0.53). In the other two cases we got same result (0.58) which is between and this for UK dataset but for other dataset maybe the result will differ. In some professions like plumber where there is big deference in P value between two cases like (Unknown male(0.1) and and unknown female(0.7)) this refer to have big number of missing values, and when we remove those value and fill gaps with as to respect the ranking we get(0.2) but if consider (Unknown 50%M+50%F) we get 0.4.



Query	Unknown male	Unknown removed	Unknown female	Unknown 50%M+50%F
architect	0.7	0.7	0.7	0.7
chef	0.4	0.4	0.6	0.5
consultant	0.6	0.6	0.6	0.6
dentist	0.5	0.5	0.6	0.5
designer	0.6	0.7	0.7	0.7
developer	0.5	0.6	0.6	0.5
doctor	0.6	0.6	0.7	0.6
economist	0.6	0.6	0.6	0.6
engineer	0.3	0.3	0.5	0.5
firefighter	0.4	0.4	0.4	0.4
gardener	0.2	0.2	0.3	0.3
hairdresser	0.8	0.8	0.9	0.8
instructor	0.7	0.7	0.8	0.8
judge	0.5	0.5	0.6	0.6
mechanic	0.2	0.2	0.2	0.2
nurse	0.8	0.9	0.9	0.9
pharmacist	0.6	0.7	0.8	0.6
photographer	0.6	0.6	0.7	0.6
physiotherapist	0.7	0.7	0.8	0.7
pilot	0.1	0.1	0.1	0.1
plumber	0.1	0.2	0.7	0.4
postman	0.2	0.2	0.3	0.2
psychiatrist	0.4	0.6	0.7	0.6
psychologist	0.9	0.98	0.98	0.98
radiographer	0.7	0.8	0.9	0.7
recruiter	0.9	0.9	0.9	0.9
reporter	0.6	0.7	0.7	0.6
salesman	0.2	0.2	0.4	0.4
scientist	0.4	0.5	0.6	0.5
secretary	0.7	0.7	0.8	0.7
surveyor	0.7	0.8	0.8	0.7
teacher	0.7	0.7	0.7	0.7
veterinarian	0.6	0.8	0.9	0.7
writer	0.6	0.7	0.8	0.6
Avarage	0.53	0,58	0,66	0.58

Figure 1: Maximum value of p retrieved by the FA*IR algorithm considering as protected grouped the female gender, we show the four cases for handling missing values in gender column in Uk linkedin, k=15 for all situations.

3 The Notion of More Protected

In this section we study the status of being more protected with respect to others attributes. In our case we consider the situation of protected group the gender and the number of star. we perform this study on doctor dataset and we consider in each country four situations. First we consider the protected group the gender female when number of star equal to 0, gender female when number

Query	p = True			
	$numstar > 0$ & female	$numstar = 0$ &female	$numstar > 0$ &male	$numstar = 0$ & male
cirugia ora	0.2	0.2	0.5	0.5
dermatologia	0.8	0.1	0.3	0.1
ginecologia	0.3	0.1	0.6	0.3
neurocirugia	0.1	0.1	0.8	0.4
oftalmologia	0.6	0.1	0.4	0.3
otorrinolar	0.7	0.1	0.4	0.1
pediatria	0.1	0.4	0.5	0.4
psicologia	0.7	0.2	0.2	0.4
psiquiatria	0.3	0.1	0.6	0.3
urologia	0.1	0.1	0.8	0.3
Avarage	0.39	0.15	0.51	0.31

Table 3: Mexico, Maximum value of p retrieved by the FA*IR algorithm considering as protected grouped the gender and number of stars in different cases .

of star more than 0, gender male when number of star equal to 0, and finally gender male when number of star more than 0.
We have two main observation, First in Mexico and Spain there is big difference in average of p value between doctors (male or female with number of star more than 0) doctors (male or female with number of star equal to 0), we can notice that female with number of star more than 0 are more protected than female

Query	p = True			
	$numstar > 0$ & female	$numstar = 0$ &female	$numstar > 0$ &male	$numstar = 0$ & male
cirugia ora	0.1	0.3	0.3	0.7
dermatologia	0.8	0.2	0.2	0.2
ginecologia	0.2	0.2	0.7	0.3
neurocirugia	0.1	0.1	0.4	0.4
oftalmologia	0.3	0.3	0.7	0.2
otorrinolar	0.1	0.4	0.6	0.2
pediatria	0.3	0.4	0.4	0.4
psicologia	0.4	0.5	0.3	0.2
psiquiatria	0.3	0.4	0.3	0.3
urologia	0.1	0.1	0.9	0.1
Avarage	0.27	0.29	0.48	0.3

Table 4: Colombia, Maximum value of p retrieved by the FA*IR algorithm considering as protected grouped the gender and number of stars in different cases .

Query	p = True			
	$numstar > 0$ & female	$numstar = 0$ &female	$numstar > 0$ &male	$numstar = 0$ & male
cirugia ora	0.1	0.1	0.7	0.5
dermatologia	0.6	0.1	0.6	0.1
ginecologia	0.3	0.1	0.8	0.1
neurocirugia	0.2	0.1	0.8	0.2
oftalmologia	0.2	0.2	0.8	0.1
otorrinolar	0.2	0.1	0.8	0.3
pediatria	0.3	0.2	0.6	0.1
psicologia	0.5	0.5	0.3	0.1
psiquiatria	0.3	0.2	0.5	0.3
urologia	0.1	0.1	0.9	0.2
Avarage	0.28	0.17	0.68	0.2

Table 5: Spain, Maximum value of p retrieved by the FA*IR algorithm considering as protected grouped the gender and number of stars in different cases.

with number of star equal to 0, in both Spain and Mexico. We can also observe that male gender with number of star equal to 0 is has average p value bigger than female with number of star equal to 0

4 Exposure Concepts

There are now substantial arguments and precedent that many of the ranking systems in use today have responsibility not only to their users, but also to the items that had being ranked. In particular, the scarce resource that ranking systems allocate is the exposure of items to users, and exposure is largely determined by position in the ranking and so is a job applicant's chances to be interviewed by an employer, an AirBnB host's ability to rent out their property, or a writer to be read. This exposes companies operating with sensitive data to legal and reputation risks, and disagreements about a fair allocation of exposure have already led to high-profile legal challenges .It is unlikely that there will be a universal definition of fairness that is appropriate across all applications, but here we will discuss two concrete examples where a ranking system may be perceived as unfair or biased in its treatment of the items that had being ranked, and where the ranking system may want to impose fairness constraints that guarantee some notion of fairness. In both examples We consider a standard exposure drop-off (i.e., position bias),

$$\frac{1}{\log_2(1+j)} \tag{1}$$

Where j is the position in the ranking, as commonly used in the Discounted Cumulative Gain (DCG) measure.

4.1 UK Linked-in

In this experiments we will try to illustrate how fairness can be related to a biased allocation of opportunity, misrepresentation of real-world distributions, and fairness. We study the linkedin UK dataset, we consider the top 15 elements in the dataset, we consider the relevance is experience level, and we assign number to each experience level as following, entry-level= 70%, mid-level=75%, senior1-level=80%, senior2-level=85%. Then for each gender we compute the average relevance in each job query, and also we compute the average exposure based on the gender position in the ranking list based on the equation given in 1 and we got the results shown in image, we can see in professions like(psychiatrist) when the average relevance male overcome the female average relevance in Small difference (0.01) it can lead to a large difference in exposure (an opportunity) for the group of females (0.18, 0.15) and even when there is no difference (pilot, pharmacist, veterinarian) we got the large difference in exposure (0.39, 0.08, 0.21), and even in some cases like (engineering) where the female average relevance overcome male average relevance (0.01) we got the large difference in exposure (0.18) for the group of females.

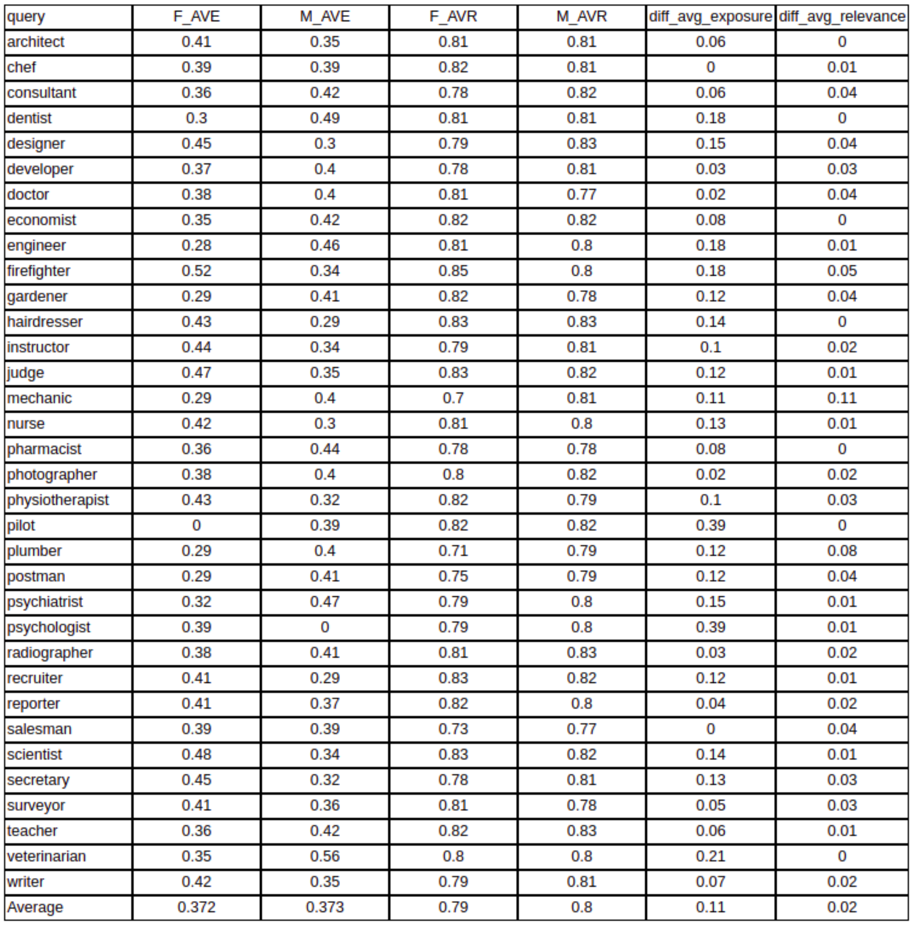
4.2 Colombia Doctors

In this example we study difference in relevance and difference in exposure in Colombia doctors dataset. We consider the top 20 elements in the dataset, and we represent the relevance by the number of star, and also we compute the average exposure based on the gender position in the ranking list based on the equation given in 1 and we got the results shown in table 6, we can notice from the table, in general there is no big deference between average exposure and average relevance, some specialist like (cirugia oral, maxilofacial and psicologia) the average relevance of female bigger than the male one, but the average exposure of male bigger than female one, but also in some cases it happen the opposite case where male average relevance overcome the female average relevance, and even though, the female average exposure overcome the male one like in (endocrinologia and otorrinolaringologia).

References

[1] Sara Galindo Martínez. Evaluating potential biases in commercial people search engines, 2019.

[2] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. *ArXiv*, abs/1706.06368, 2017.



query	F_AVE	M_AVE	F_AVR	M_AVR	diff_avg_exposure	diff_avg_relevance
architect	0.41	0.35	0.81	0.81	0.06	0
chef	0.39	0.39	0.82	0.81	0	0.01
consultant	0.36	0.42	0.78	0.82	0.06	0.04
dentist	0.3	0.49	0.81	0.81	0.18	0
designer	0.45	0.3	0.79	0.83	0.15	0.04
developer	0.37	0.4	0.78	0.81	0.03	0.03
doctor	0.38	0.4	0.81	0.77	0.02	0.04
economist	0.35	0.42	0.82	0.82	0.08	0
engineer	0.28	0.46	0.81	0.8	0.18	0.01
firefighter	0.52	0.34	0.85	0.8	0.18	0.05
gardener	0.29	0.41	0.82	0.78	0.12	0.04
hairdresser	0.43	0.29	0.83	0.83	0.14	0
instructor	0.44	0.34	0.79	0.81	0.1	0.02
judge	0.47	0.35	0.83	0.82	0.12	0.01
mechanic	0.29	0.4	0.7	0.81	0.11	0.11
nurse	0.42	0.3	0.81	0.8	0.13	0.01
pharmacist	0.36	0.44	0.78	0.78	0.08	0
photographer	0.38	0.4	0.8	0.82	0.02	0.02
physiotherapist	0.43	0.32	0.82	0.79	0.1	0.03
pilot	0	0.39	0.82	0.82	0.39	0
plumber	0.29	0.4	0.71	0.79	0.12	0.08
postman	0.29	0.41	0.75	0.79	0.12	0.04
psychiatrist	0.32	0.47	0.79	0.8	0.15	0.01
psychologist	0.39	0	0.79	0.8	0.39	0.01
radiographer	0.38	0.41	0.81	0.83	0.03	0.02
recruiter	0.41	0.29	0.83	0.82	0.12	0.01
reporter	0.41	0.37	0.82	0.8	0.04	0.02
salesman	0.39	0.39	0.73	0.77	0	0.04
scientist	0.48	0.34	0.83	0.82	0.14	0.01
secretary	0.45	0.32	0.78	0.81	0.13	0.03
surveyor	0.41	0.36	0.81	0.78	0.05	0.03
teacher	0.36	0.42	0.82	0.83	0.06	0.01
veterinarian	0.35	0.56	0.8	0.8	0.21	0
writer	0.42	0.35	0.79	0.81	0.07	0.02
Average	0.372	0.373	0.79	0.8	0.11	0.02

Figure 2: Linkedin in United Kingdom, example to illustrate difference in relevance and difference in exposure for the group of females and males.

	F AVE	M AVE	F AVR	M AVR	diff avg expo	diff avg relev
angiologia y cirugia vasc	0	0.35	0	1.5	0.35	1.5
cardiologia adultos	0.62	0.32	0	2.17	0.3	2.17
cirugia general	0.5	0.34	5	2.63	0.16	2.37
cirugia oral y maxilofacial	0.3	0.37	1.25	0.94	0.07	0.31
cirugia plastica este y repa	0.37	0.35	1.25	2.19	0.02	0.94
dermatologia	0.36	0.32	3.94	2.5	0.04	1.44
endocrinologia	0.44	0.31	0.83	2	0.13	1.17
gastroenterologia especial	0.28	0.38	3	3.07	0.1	0.07
ginecologia y obstetricia	0.29	0.37	2.8	3.67	0.08	0.87
medicina estetica	0.29	0.5	2.14	1.17	0.21	0.97
neurocirugia	0.32	0.35	5	1.32	0.03	3.68
odontologia y estomatol	0.31	0.4	2.27	2.22	0.09	0.05
oftalmologia	0.35	0.35	2.71	3.77	0	1.06
otorrinolaringologia	0.49	0.29	0.83	3.14	0.2	2.31
pediatria	0.37	0.33	1.4	2.5	0.04	1.1
psicologia	0.32	0.45	1.93	3	0.13	1.07
psiquiatria	0.42	0.3	1.56	1.82	0.12	0.26
reumatologia	0.43	0.31	0.71	1.92	0.12	1.21
traumatologia y ortoped	0.3	0.36	4.5	3.44	0.06	1.06
urologia	0	0.35	4.5	4.35	0.35	0.15
Avarage	0.338	0.355	2.281	2.466	0.13	1.188

Table 6: Colombia doctors example, illustrate difference in relevance and difference in exposure for the group of females and males.