Biases in Machine Learning

Nour Eldin Badr[1]

Bielefeld University, Bielefeld, Germany

Abstract. While training machine learning models you can find that your model is biased, this bias can race bias, gender bias or any other bias, this bias can affect many critical decisions can be taken like in court in predicting re offended or deciding if we should give a loan or not to a certain person, These biases is important to detect it, measure it and overcome it in our models.

Keywords: Biases · Race Bias · Gender Bias · machine Learning .

1 Introduction

In the last few years machine learning and artificial intelligence has become one of the huge technological trends around the world. Country's leaders and scientists are trying to merge it on our daily life, for example, nowadays machine learning could help in taking some decisions such as approving a loan from a bank or even choosing qualified applicants for an interview or preparing for an interview as a chatbot as in figure 1 .But here comes a small problem that this decision could biased, then it would be a problem because it affects honesty in taking decisions,this bias can be caused from the data we are feeding to the machine learning algorithms which we got from the people themselves who might have conscious or unconscious preferences.In Fact we cant only blame the people who we got from them the data, but some problems could occur from researchers who have lacked in some cognitive assessments. From here scientists and organizations who are working with gathering data sets and working with machine learning have agreed that they have to find as a solution for this problem. From my opinion, I think them to check the data that they are going to use it to train their models,They should represent the data with different representations like genders, races, cultural representations to develop the algorithms that would have an effect on the data samples and will shape it to decrease biases.In this essay we will have around on two different biases which are gender and racial biases, and how can we fight these approaches.

2 Gender and Race Bias

Before going through this section and explain gender and race bias, we would like to give first an overview about what exactly biases is. In cognitive science it is defined in general as the deviation from some true or objective value or

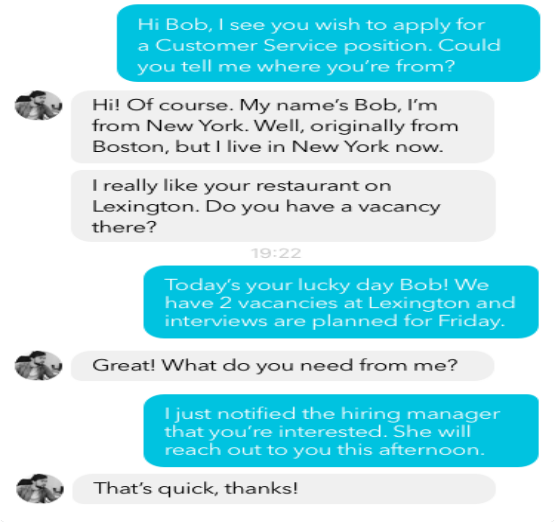

Fig. 1. chatbot helping in recruitment[1]

we can say the deviation from a norm.These biases can be easy to identify and sometimes it can be difficult to identify, all depends on the normative model, if the normative model is describing how the bass should be present then it is easy to identify while if it is unknown then it is difficult to identify [6].Some scientists like Goldstein and Gingerenze has assumed that human brain always tends to use less cognitive efforts so the brain applies shortcuts with trying to process the whole information, but part of it, so here bias will appear because of the lake of the full information[7].As now we have known what is a bias, we can discuss about gender bias and race bias.

2.1 Gender Biases

In this bias the machine is always biased toward one of the genders either male or female because of these bias people might find inequality in occupations and in job interviews. According to this paper, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings [5],They are claiming that the problem is in word embedding, word embedding in general is one of the representations of document vocabulary or we can say it is a vector representation of a specific word, we are using it to know similar words like queen and king, greets and speaks and to state the distance between them as in figure 2 [8]. Let us have an example to understand it better.If you want to search about PhD students who are studying informatics at Bielefeld university, you will find that the word embedding is trying to the terms that are near or related to informatics closer to male names rather than the female names (e.g., the embeds give Mark: Computer Programmer while it gave Sally: house-keeper). So what

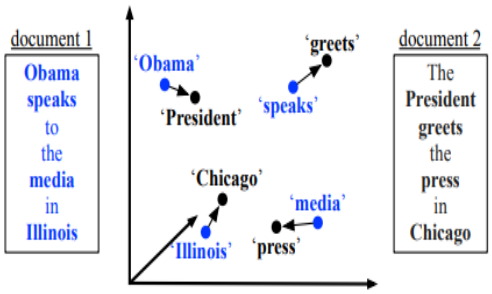

Fig. 2. Word Embedding Example[8]

we will notice that the word embedding affected the searching engine to give Mark a rank higher than sally and finally you will find the searching engine will give you firstly the male computer scientist at Bielefeld than the females at the end, and this actually occurred, From this we conclude that word embedding has made it tougher for woman to be recognized as computer scientist[5]. Figure 3 shows how the occupations is being classified between males and females.

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Fig. 3. gender bias example in occupations from g2vNews[5]

One of the examples also of gender bias has appeared in google translation when you try to translate in the Google translation from Turkish to English like the example shown in figure 4, you will find that doctor is being classified as male and nurse will be classified as female.Finally, according to Man is to Computer Programmer as Woman is to Homemaker? Debasing Word Embeddings [5] paper, they have said that these problems we have discussed is happening because word embeddings need huge data to extract associations and relationships as they are trained by some methods like second order methods [5].Finally we have seen how can the gender bias affect the occupation and also the translation .

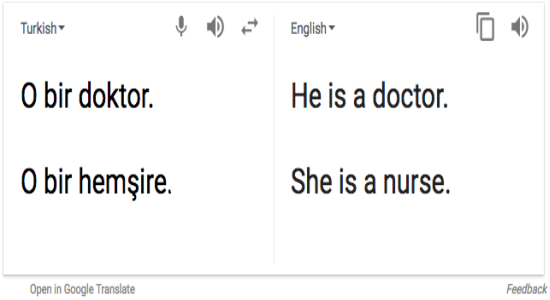

Fig. 4. gender bias example in google translation[2]

2.2 Race Biases

While using machine learning and applying it in different cases in our daily life, scientists have found that machine learning algorithms could produce Race bias which is sometime really dangerous for taking some actions or decisions. In this part we are going to give two examples of racial bias.
You can find race bias in ranking the risk of criminals or what is the probability of the criminal to be reminded again. An organization called propublica has analyzed an AI program called COMPAS which has been used to predict these probabilities [9]. Criminals take some a survey of 137questions and then the software will try to predict how likely, the defendant is to re-offend based on his answers to this survey. What this organization has found, is that black criminals will be reoffended with high risk more than white criminals [9].Despite the offenses of the white criminals with low risk was more dangerous than black criminals with high risk, But COMPAS has assigned high risk to black criminals as shown in figure 5.

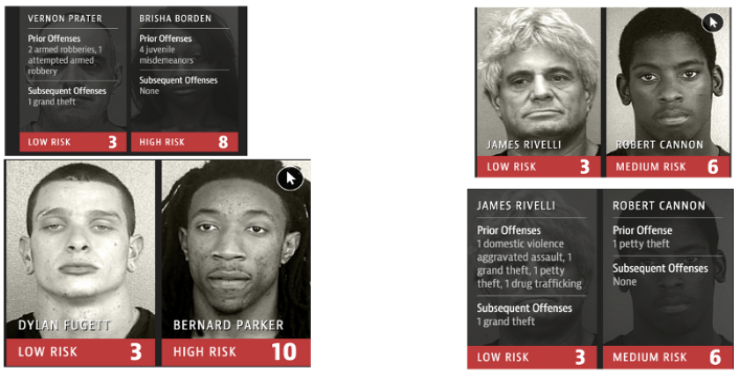

Fig. 5. COMPASS has assigned high risk to black criminals and low risk to white criminals [3]

Another example of racial bias can appear in detecting the dialect of African-American English in social media as it is different from American English.In this paper "Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English" [4].They have analyzed an African-American English Twitter corpus and what they have found racial inequality in language identification.Then they had a small experiments on different NLP platforms, but at this time they have increased the message length that can be written.What they have found that when you increase the tweet length, the classification would be more accurate and the race disparity will decrease between the African American tweets and white American tweets [4]. the following figure 6 shows how the accuracy decrease as the message length increase within different platforms.Finally race bias from my opinion is really important and critical as it can affect some critical decisions also it is missing in one of the sensitive topics around the world .

		AA Acc.	WH Acc.	Diff.
langid.py	$t \leq 5$	68.0	70.8	2.8
	$5 < t \leq 10$	84.6	91.6	7.0
	$10 < t \leq 15$	93.0	98.0	5.0
	$t > 15$	96.2	99.8	3.6
IBM Watson	$t \leq 5$	62.8	77.9	15.1
	$5 < t \leq 10$	91.9	95.7	3.8
	$10 < t \leq 15$	96.4	99.0	2.6
	$t > 15$	98.0	99.6	1.6
Microsoft Azure	$t \leq 5$	87.6	94.2	6.6
	$5 < t \leq 10$	98.5	99.6	1.1
	$10 < t \leq 15$	99.6	99.9	0.3
	$t > 15$	99.5	99.9	0.4
Twitter	$t \leq 5$	54.0	73.7	19.7
	$5 < t \leq 10$	87.5	91.5	4.0
	$10 < t \leq 15$	95.7	96.0	0.3
	$t > 15$	98.5	95.1	-3.0

Fig. 6. Percent of the 2,500 tweets in each bin classified as English by each classifier : Diff. this the difference (disparity on an absolute scale) between the classifier accuracy on the AA aligned and white aligned samples. t is the message length for the bin[4]

3 How to overcome Race and Gender biases

In this section we are going to explain the different solutions that can overcome biases we have discussed in artificial intelligence and machine learning like race bias and gender bias.According to "Ethics and Bias in Machine Learning: A Technical Study of What Makes Us Good "[9] paper they have clarified some solutions like ethical, technical ,political and social solutions.

3.1 Ethical solution

From my opinion, i would say that the ethical solution will come firstly from the data, scientists who are dealing with data, I would say they shouldn't feed

model with data which is already biased too much, they should filter that data from biases and they have to realize that most of the data are biased because we live in a prejudiced world.

3.2 Technical solution

In the technical solution form my opinion, it is really important also for the machine while training the model to determine which is ethical and which is not. I would also say we can make machine learning models that can predict racial bias and gender bias to overcome it and to know how much bias, we can find in our data and what is the source of it. One of the solutions also I think we can measure the bias in our model then subtract it.That was my opinion, but "Ethics and Bias in Machine Learning: A Technical Study of What Makes Us Good " [9] paper has another opinion, they are thinking of using the Markov model Tree to decrease variance which they have related it to statistical bias as they are trying to determine the source of bias , through this tree there will be data on the split and the leaves of the tree and they will use multiple hypothesis to vote on the classification of the test cases through randomization methods [9]. this will lead to less errors and biases in test cases.

3.3 Social solution

From my opinion social solutions are the most important solution among the solutions we have discussed, as we are collecting data which may can contain biased and prejudiced data from the society.Just imagine if we are training a Chabot1 on some data from users who are using bad words and slangs and the other Chatbot2 is trained on polite conversations,you will find Chatbot1 is cursing the users after going to the market while Chatbot2 will have a great success among the community as it is polite unless someone again is teaching it biases.So it is important to have social awareness about machine learning, how it is used and why.

3.4 Political solution

Machine learning now a days can be used in politics and politicians can manipulate people through machine learning, especially in elections by the data they are collecting from social media and then knowing which region supporting whom and they can take decisions what to do there, as we have seen also in the last section how machine learning is being used in predicting the risk percentage of the criminals. So I think that politicians should be sure that the machine learning is not biased, thats why they should apply some standards, regulations, policies and retraining certifications to machine learning sessions. And I think this happened in EU and US,when the EU passed a regulation which outlines citizens right to an explanation regarding machine learning decisions made about them [9], while in the US Obama administration tried to ensure fairness, thats why they have tried to push investigation for big data and machine learning algorithms [9].

4 Conclusion

While training machine learning models you can find that your model is biased, this bias can race bias, gender bias or any other bias, this bias can affect many critical decisions can be taken like in court in predicting re offended or deciding if we should give a loan or not to a certain person, these biases is important to detect it, measure it and overcome it in our models.

References

1. chatbotmessage, https://harver.com/blog/uses-ai-in-recruitment/
2. googletranslation, https://translate.google.com/
3. Racebias, https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1
4. Blodgett, S.L., O'Connor, B.: Racial disparity in natural language processing: A case study of social media african-american english. arXiv preprint arXiv:1707.00061 (2017)
5. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in neural information processing systems. pp. 4349–4357 (2016)
6. Caverni, J.P., Fabre, J.M., Gonzalez, M.: Cognitive biases: Their contribution for understanding human cognitive processes. In: Advances in psychology, vol. 68, pp. 7–12. Elsevier (1990)
7. Goldstein, D.G., Gigerenzer, G.: " models of ecological rationality: The recognition heuristic": Clarification on goldstein and gigerenzer (2002). (2002)
8. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International Conference on Machine Learning. pp. 957–966 (2015)
9. Shadowen, A.N.: Ethics and bias in machine learning: A technical study of what makes us good (2017)