Final Project

"Team Y"- Yuval Segal & Eran Aizikovich

"What is the connection between an annotator's background and the way he will rate hate speech?"

Our GitHub Repository - https://github.com/yo21/algorithm-bias-R

Data Source - https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

Section 1 - Introduction

We live in an information age in which many decisions are made by various software programs that use data annotated by humans. These types of decision-making are seemingly sterile and free from prejudice. That perspective gives software decision-making a huge public validation since software programs are unbiased, but is that so?

In this project, we decided to focus on the field that is known as "Algorithmic bias". The reason we chose this subject is that it has an enormous impact on us as a society daily.
The ambition to reach equality is critical in many fields where wrong decision-making could be catastrophic, for example, research on Racial-Bias in Health Care, published in Science AAAS, shows that since less money is spent on black patients, the algorithm concludes they are healthier than white patients – even though this is due ongoing discrimination. Another research on the subject is Gender-Based Discrimination published in SSRN. Furthermore, the field of "Algorithmic Bias" concerns 2 out of the 17 UN official Sustainable Development Goals, Goal 5: Gender Equality & Goal 10: Reduced Inequalities.

Supervised learning models that are used in various software programs use train sets – or in other words, annotated data. The question we wish to ask is: "Could there be a significant biased connection between the annotator's background and the way they will rate hate speech[1]?".

In this project, we will try to examine the relationship and the correlation between different sectors in the society (by race, religion, age, gender, etc.) and see if there is evidence of a common denominator between a sector while annotating a hate speech post that is targeting a specific group.
We want to innovate and show you that there has to be a different way of dividing the data to annotators while considering their background – for example, should extreme conservative annotators rate data that regards hatred against their own religion. In contrast, should they annotate data that concerns hatred against liberals?

[1] Hate speech definition: "bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics, especially when the group is unnecessarily labeled."

Section 2 – Data overview

The data we used has been published in an article from Berkeley University. The data contains 131 variables, most of them were binary and for our project, we transformed them into categorical variables, alongside irrelevant (to us) variables that we ignored. The variables we focused on are the ones that were filled by the annotators.

The way the data was collected – 7912 different annotators were given posts and to every post, they were asked to rate by the following measures:

- Identities of the target group – such as race, religion, gender, sexual orientation, etc....
- Whether the comment is Hate-Speech – 0 if no, 2 if yes, and 1 when unclear.

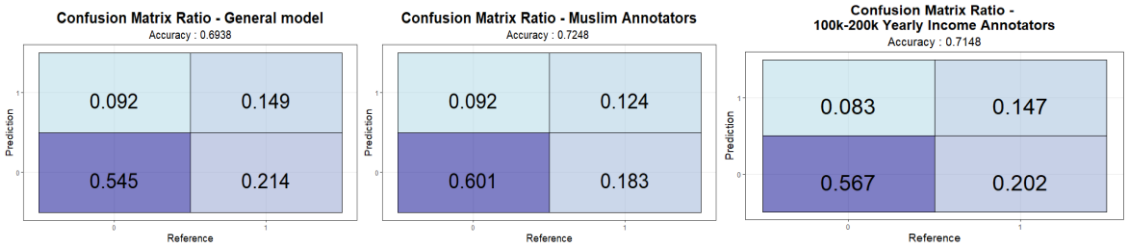
When finished annotating, the rater was asked to fill information about himself.

The Dataset we used after filtering the unnecessary variables held the following features:

1. "sentiment", "respect", "insult", "humiliate", "dehumanize", "violence", "genocide", "attack_defend". variables scaled from 0 to 5 for the comment.
2. Target: race, religion, gender, sexuality- categorial variables that specify the identity of the target in the comment.
3. Annotator: gender, education, income, ideology, age, race, religion, sexuality – categorial variables that specify the identity of the annotator of the comment.

Section 3 - Methods and results:

To understand the influence that backgrounds have on annotating data we try to examine how different people annotated data and find the common denominator between them.
To do so we fitted two types of logistic regression models that predict will a person classify a post as hate speech(Y/N): The general population model, Sectorial model – which have a common background feature.
In research to find the best features to train the models, we found that the best features are target race, target religion, target gender, and target sexuality. These models resulted in the following confusion matrix:



Key Results & Conclusions

1. By addressing the data through the different categories in the annotator's background the model predicts better than the general model
2. coefficients – From the significant coefficients, we learned that groups with different backgrounds grasp differently, posts that target specific sectors. For example, we noticed that annotators with a Ph.D. are more sensitive to comments that are targeting gays than annotators with a professional degree.

Section 4 - Limitations and Future Work:

In our approach, we had some critical limitations,

- None of the models we tried were good enough in identifying hate speech, and that led us to analyze the data without that critical verdict.
- Trying to fit regression models was frustrating because a lot of the data is categorical (or binary), and the result was not always clear to grasp.
- We believe that more sophisticated ML algorithms could have worked here better and through them, we could get better results.

With more time we hope that we could get a better understanding of the data (that by itself is exceptionally good and rich with details) and analyze it more thoroughly to achieve more meaningful insights such as: investigating the relations between more features in our data and trying to estimate the influence a person's background has when criticizing/annotating data that the target shares a mutual background. For example, a Jewish person annotating an antisemite post. Since we have seen some hints of that being meaningful but sadly did not have the time, nor the resources to investigate that direction.

Appendix

- Our GitHub: https://github.com/yo21/algorithm-bias-R

- Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application: https://arxiv.org/pdf/2009.10277.pdf

- A white mask worked better: why algorithms are not color blind. https://www.theguardian.com/technology/2017/may/28/joy-buolamwini-when-algorithms-are-racist-facial-recognition-
- How to write a racist AI in R without really trying. https://notstatschat.rbind.io/2018/09/27/how-to-write-a-racist-ai-in-r-without-really-trying/

UN official Sustainable Development Goals

- Goal 5: Gender Equality- https://www.un.org/sustainabledevelopment/gender-equality/
- Goal 10: Reduced Inequalities-https://www.un.org/sustainabledevelopment/inequality/