**Section 1 - Intro**

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. In ways of equality in many types of fields, for example, research on [Racial-Bias in Health Care](https://www.science.org/doi/abs/10.1126/science.aax2342) published in Science AAAS, another work on the subject is [Gender-Based Discrimination](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260) research published in SSRN. You may see why this is such an important subject as its concerns 2 of the 17 UN SUSTAINABLE DEVELOPMENT GOALS, [Goal 5: Gender Equality](https://www.un.org/sustainabledevelopment/gender-equality/) & [Goal 10: Reduced Inequalities](https://www.un.org/sustainabledevelopment/inequality/)

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? If so quantile its intensity”.

### what is our innovation

To try and answer this question, we used a dataset from a paper published by Berkeley University. Indeed, we found a bias in the annotating data as will explore late on in section 3. The innovation in this field is that we need to label data with more supervision to reduce this bias.

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](https://arxiv.org/pdf/2009.10277.pdf) from Berkeley University. The data contain 131 variables, most of them were binary and for our project, we transform them into categorical variables, alongside some irrelevant to us variables 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 this post as follows:

* Identities of the target group – such as race, religion, gender, sexual orientation, etc.…
* Whether the comment is Hate-Speech – 0 for not 2 for yes and 1 for unclear.

When finished annotating, the rater was asked to fill in the flowing information about himself:

* Yearly income, Gender, Political ideology, Race, Religion, Education, and Sexuality.

The Dataset we used after filtering the unnecessary variables hold

1. "sentiment", "respect", "insult", "humiliate", "dehumanize", "violence", "genocide", "attack\_defend". variables scaled from 0 to 5 for the comment.
2. "target\_race", "target\_religion", "target\_gender", "target\_sexuality", categorial variable that specify the identity of the target in the comment.
3. "annotator\_gender", "annotator\_educ", "annotator\_income", "annotator\_ideology", "annotator\_age", "annotator\_race", "annotator\_religion", "annotator\_sexuality" categorial variable - specify the identity of the annotator of the comment.
4. Hate speeches, hate\_speech\_score, std\_err

**Section 3 - Methods and results:**

To understand the influence that backgrounds have on annotating data we try exam the different people annotated data and find the common denominator between them. To do so we fitted two types of logistic regressions models,

* General population model – predict hate speech.
* Sectorial models - predict hate speech for a given sectorial background.

A picture containing diagram

Description automatically generatedA picture containing diagram

Description automatically generatedIn research to find the best features to train the models, we found that the best features are target\_race, target\_religion, target\_gender, target\_sexuality, and annotator\_age. Lets the models

A picture containing diagram

Description automatically generated

Key results –

1. it is easy to see, that by addressing the data through the different categories of the annotator’s background the models achieve a better prediction.
2. Significant coef - …

**Section 4 - Limitations and Future Work:**

In our approach we had some critical limitations,

first none of the models we tried (k-means, random forest) where 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 where not always as clear to grasp.

Our data limited us especially from the aspect that it was time consuming to clean and decide what is relevant for us.

With more time we hope that we could get 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.

We believe that more sophisticated ML algorithms could have worked here better and through them we could gotten to better results.

In specific we would have wanted to investigate the relations between more features in our data and to try and 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) because we have seen some hints of that being meaningful but sadly did not have the time not the resources to investigate that direction.

Appendix