Project proposal

Team Y

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. For example, software for criminal sentencing was found to be biased against blacks.

Attached to this introduction you will find articles on the subject. 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".

To try and answer this question, we will use statistical methods on a dataset from a paper published by Berkeley University, we will extend more about it in Section 2.

2. Data

In this project, we are using data published in an article made in Berkeley university as explained in section 1. link to both article and dataset are attached at the end of this section. A glimpse to the data is in the appendix part.

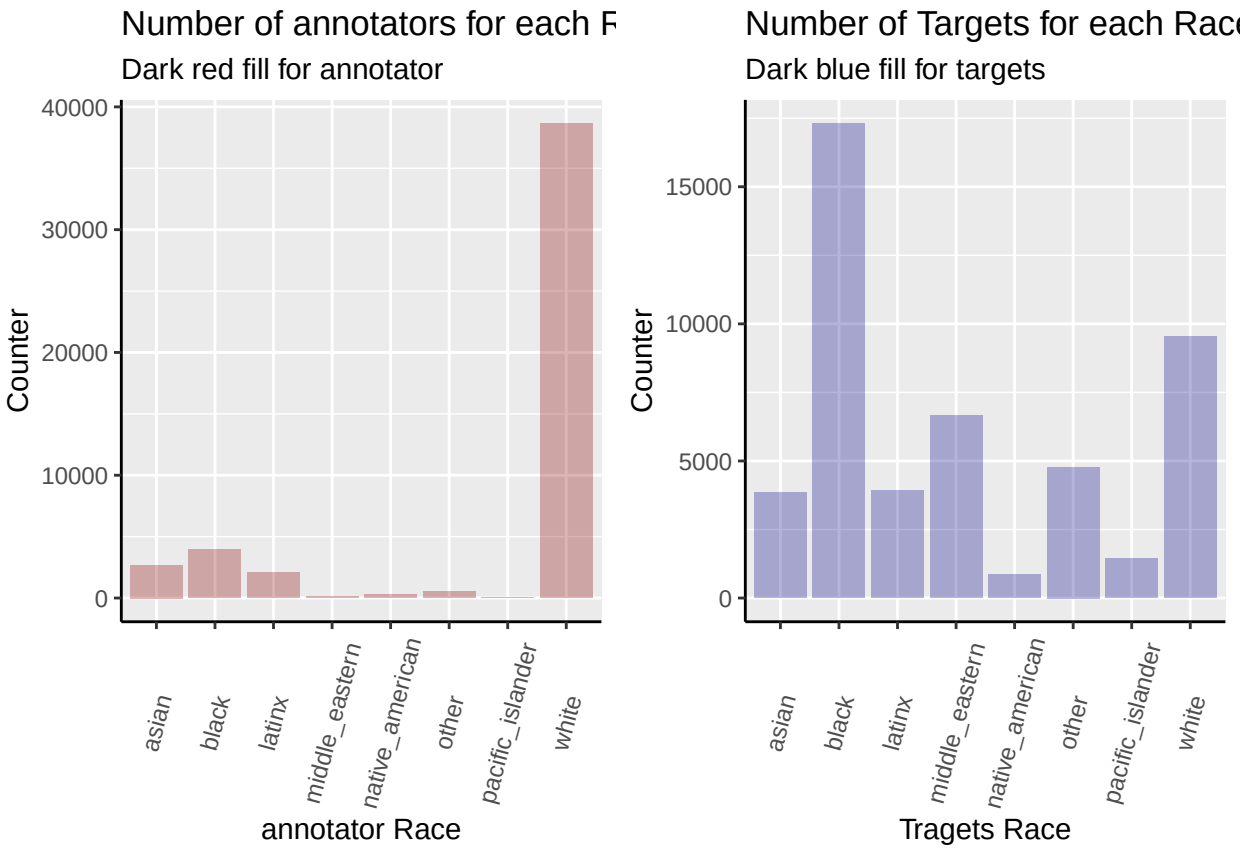
The original data contains 131 variables and 135,556 rows. To be able to work with it, we filtered it down to 30 relevant variables. the remain data holds information both on the annotator and on the post itself, we will be referencing the post as the "target" from here on.
In this data, you will find many types of variables, for example the "hate_speech_score" variable is a numerical variable, while the "annotator_race" variable is a categorical one.

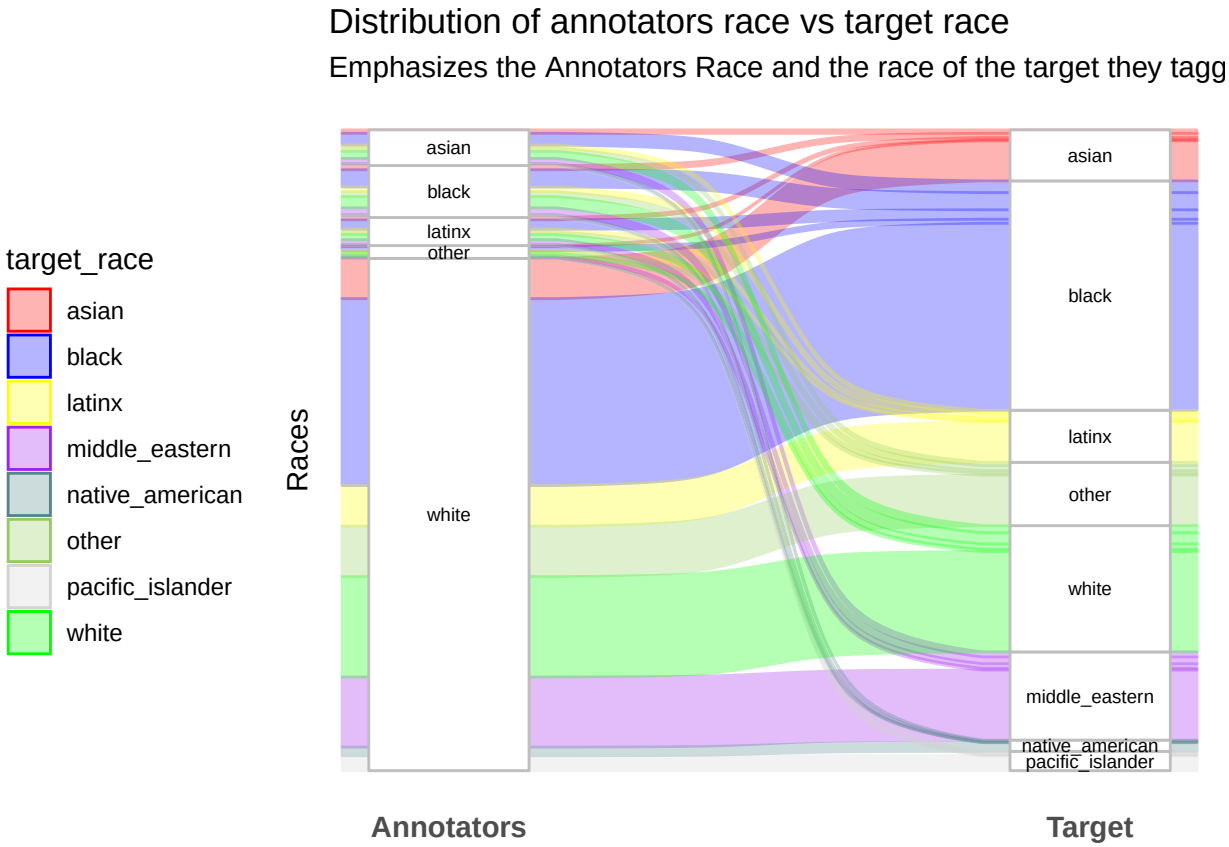
The article: https://arxiv.org/abs/2009.10277 The data: https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

All along this presntation we will mark annotator as red and target as blue.

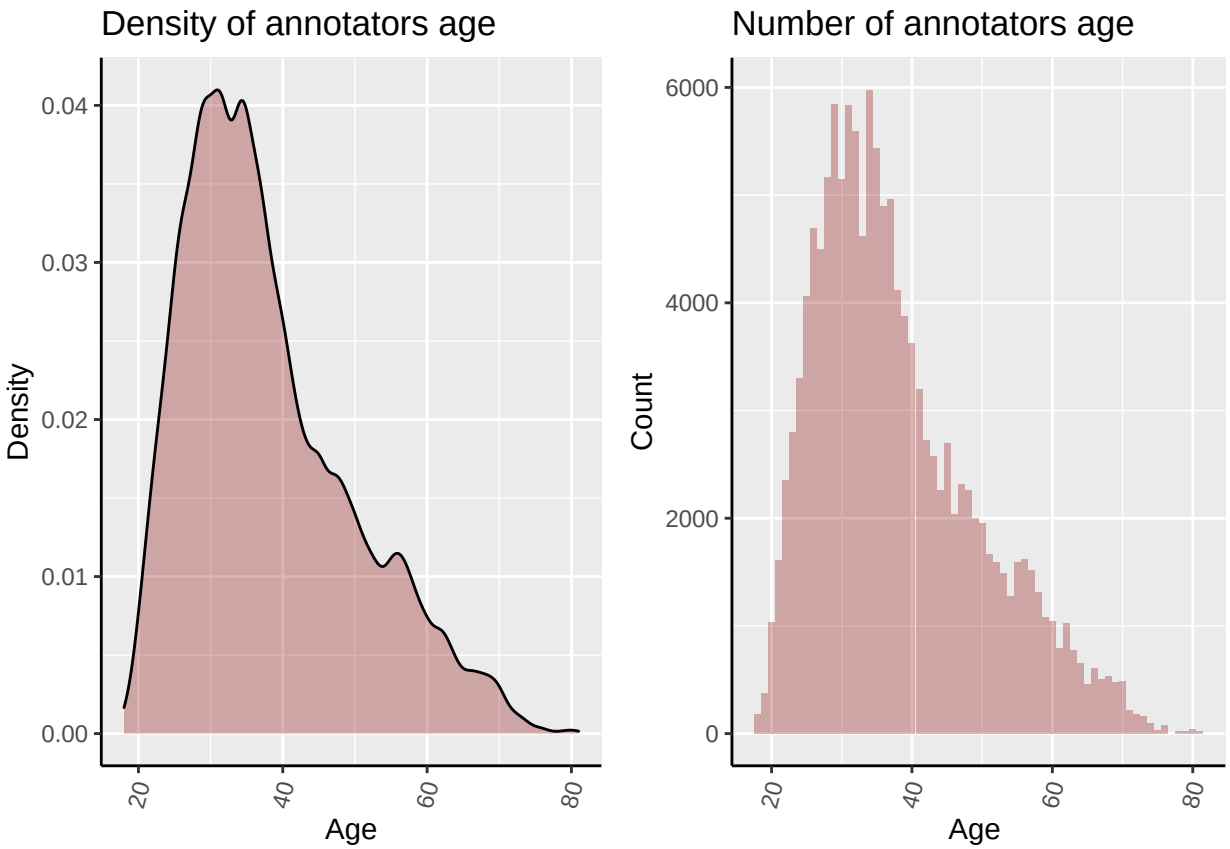
3. Preliminary results

As we can in the graph below there is a big difference between the annotators and the targets distributions. Clearly, most of the annotators are self defined as whites while most of the targets are tagged as blacks.

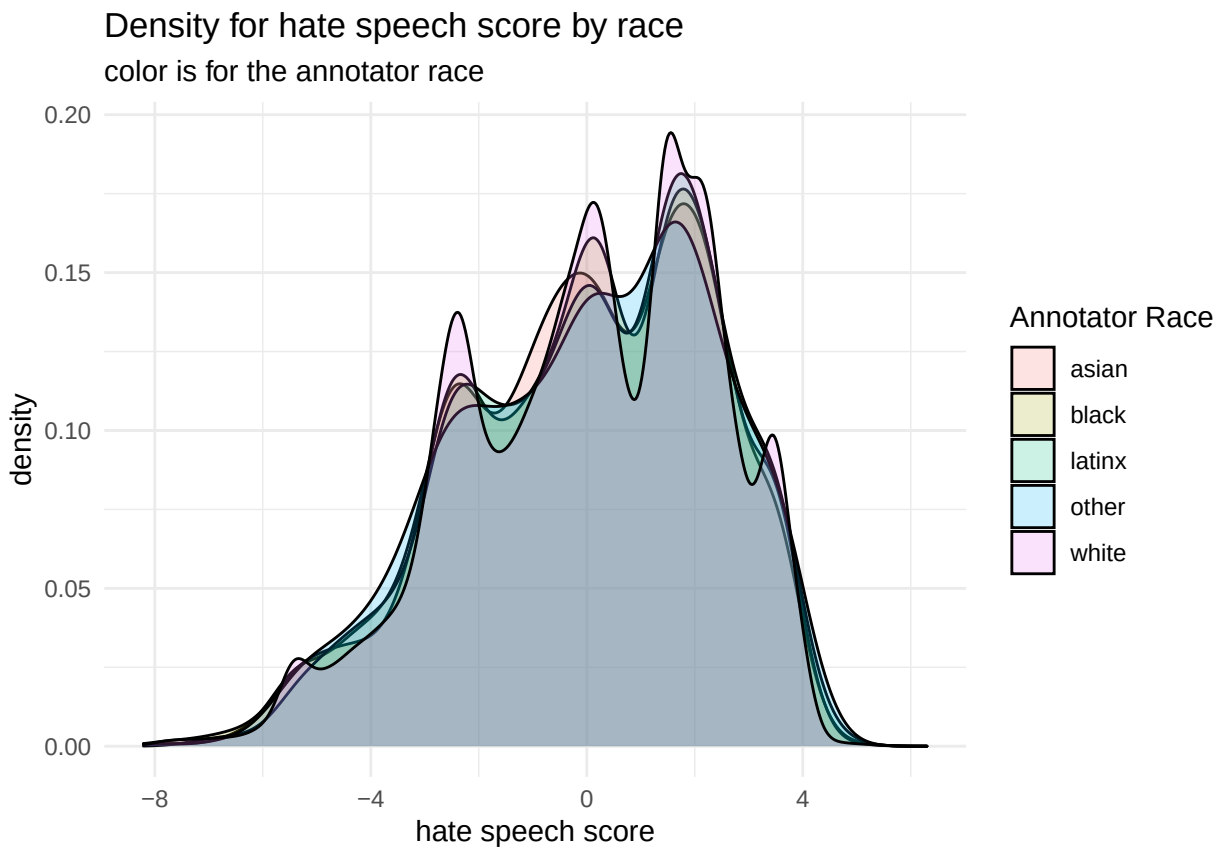

Number of annotators for each F
Dark red fill for annotator
Counter
40000
30000
20000
10000
0
asian
black
latinx
middle_eastern
native_american
other
pacific_islander
white
annotator Race
Number of Targets for each Rac
Dark blue fill for targets
Counter
15000
10000
5000
0
asian
black
latinx
middle_eastern
native_american
other
pacific_islander
white
Tragets Race



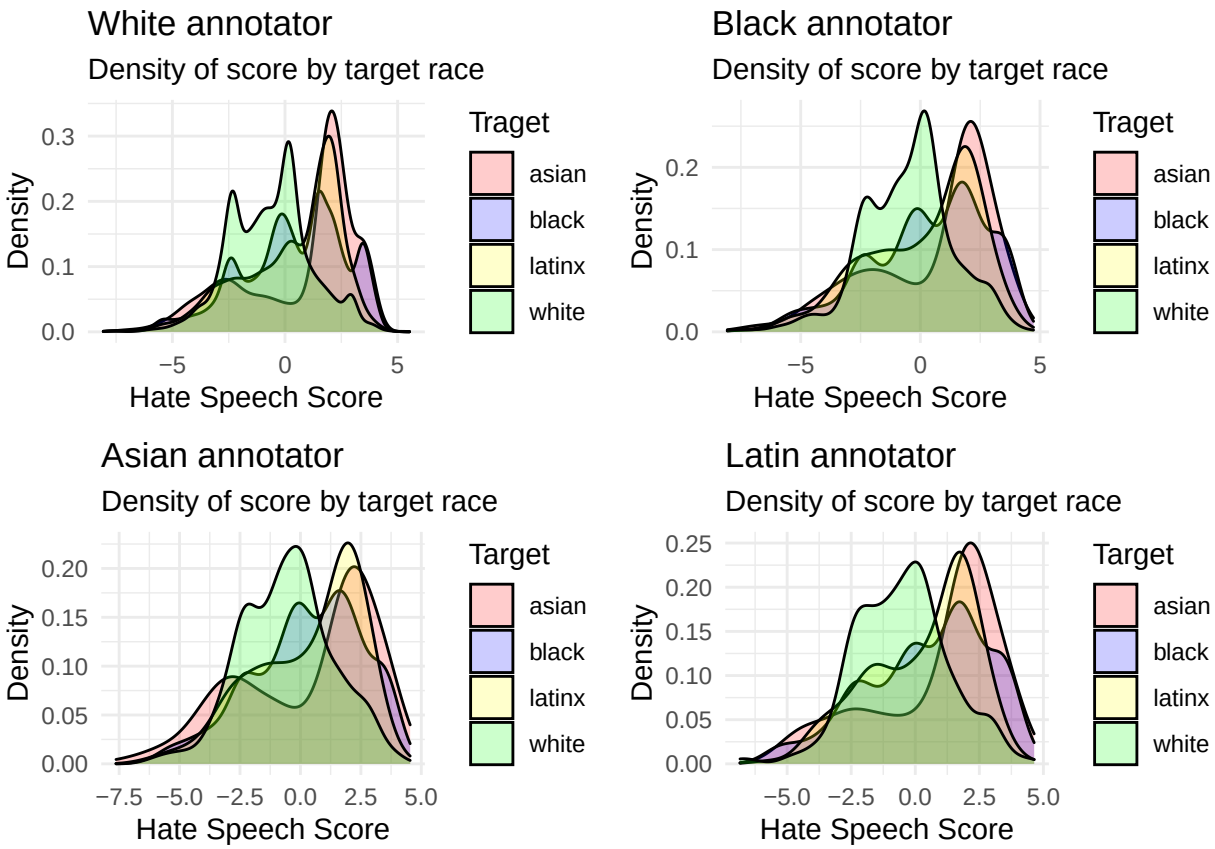
As we can see the average age of the annotator is about 30, and the distribution is inclined towards the younger age.



Here it looks like the way that all the races measure hate speech is the same but is it so? In the next graphs, we will see that not so much.



As promised in the next graphs we are diving into more detail, and it is very interesting to see that all races measured lower hate scores when their target is white.



4. Data analysis plan

We will try to predict the significance of the difference between the way annotators from different backgrounds are choosing to score hate speeches. To do so we will try to find which features gives us the largest variance and significance between different backgrounds.

For the prediction we will use basic statistical methods such as Pearson correlation, the covariance between variables, and linear regression models.

The results we would like to see, to support our hypothesis, are strong connections between the way people see hate speech and their background. We chose to divide the work as follow: one of us is more responsible for the statistical concepts while the other is more focused on preparing the data and the visualization parts.

Appendix

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

Articles about Algorithms bias:

A white mask worked better': why algorithms are not colour blind.

`https://www.theguardian.com/technology/2017/may/28/joy-buolamwini-when-algorithms-are-racist-facial-rec`

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/`

Data README

```
cat(readLines('../data/README.md'), sep = '\n')
```

```
# algorithm-bias-R
Searching for proof that algorithmic bias is significant
# SISE2601 Project data description
================
Eran Aizikovich and Yuval Segal

Our original Data set has 131 columns, most of them were redundant.
After cleaning the data, without losing any information that is meaningful
to our research we were left with 30 columns.

Rows: 135,556
Columns: 30
Primary key in our data:
    * comment_id <int>   unique ID for each comment
    * annotator_id <int> unique ID for each annotator
* sentiment <dbl>  ordinal label that is combined into the continuous score
* respect <dbl> ordinal label that is combined into the continuous score
* insult <dbl> ordinal label that is combined into the continuous score
* humiliate <dbl> ordinal label that is combined into the continuous score
* dehumanize <dbl> ordinal label that is combined into the continuous score
* violence <dbl> ordinal label that is combined into the continuous score
* genocide <dbl> ordinal label that is combined into the continuous score
* attack_defend <dbl> ordinal label that is combined into the continuous score
* hatespeech <dbl> ordinal label that is combined into the continuous score.
* hate_speech_score <dbl> The continuous score of the post, higher = more hateful .
* infitms <dbl> inlier sensitive value(continuous) of the comment.
* outfitms <dbl> outlier sensitive value(continuous) of a comment.
* annotator_severity <dbl> continuous value that estimates the annotatorâ€s survey interpretation bias.
* std_err <dbl> continuous value of the standard deviation between annotators on this comment.
* annotator_infitms  <dbl> inlier sensitive value(continuous) of the annotator.
* annotator_outfitms <dbl> outlier sensitive value(continuous) of a comment.
* target_race <chr>  categorical variable, tagged by annotator
* target_gender <chr>  categorical variable, tagged by annotator
* target_sexuality <chr> categorical variable, tagged by annotator
* annotator_gender <chr> categorical variable
* annotator_educ <chr> categorical variable, annotatorâ€s education.
* annotator_income <chr> categorical variable, within range.
* annotator_ideology <chr> categorical variable: neutral, slightly_conservative, extremely_liberal,
    extremely_conservative, liberal, conservative, slightly_liberal, no_opinion
* annotator_age <dbl> Discrete variable
* annotator_race <chr> Categorical variable
* annotator_religion  <chr> Categorical variable
* annotator_sexuality <chr> Categorical variable: straight, bisexual, gay, other


Our original data had 131 columns; the additional columns were removed by us in the cleaning stage.
The columns we removed were:
(All columns besides the first three were Boolean variables.)
```

```
platform, status, text, target_race_asian, target_race_black, target_race_latinx,
target_race_middle_eastern, target_race_native_american, target_race_pacific_islander,
target_race_white, target_race_other, target_religion_atheist, target_religion_buddhist,
target_religion_christian,   target_religion_hindu, target_religion_jewish,
target_religion_mormon, target_religion_muslim, target_religion_other,
target_origin_immigrant, target_origin_migrant_worker , target_origin_specific_country,
target_origin_undocumented, target_origin_other, target_gender_men, target_gender_non_binary,
target_gender_transgender_men, target_gender_transgender_unspecified,
target_gender_transgender_women, target_gender_women, target_gender_other,
target_sexuality_bisexual, target_sexuality_gay , target_sexuality_lesbian,
target_sexuality_straight, target_sexuality_other, target_age_children , target_age_teenagers,
target_age_young_adults , target_age_middle_aged, target_age_seniors, target_age_other,
target_disability_physical, target_disability_cognitive, target_disability_neurological,
target_disability_visually_impaired, target_disability_hearing_impaired,
target_disability_unspecific, target_disability_other , target_disability, annotator_trans,
annotator_gender_men, annotator_gender_women, annotator_gender_non_binary,
annotator_gender_prefer_not_to_say, annotator_gender_self_describe, annotator_transgender,
annotator_cisgender, annotator_transgender_prefer_not_to_say, annotator_education_some_high_school,
annotator_education_high_school_grad , annotator_education_some_college, annotator_education_college_gra
annotator_education_college_grad_ba, annotator_education_professional_degree , annotator_education_maste
annotator_education_phd , annotator_income_<10k, annotator_income_10k>50k, annotator_income_50k>100k,
annotator_income_100k>200k, annotator_income_>200k, annotator_ideology_extremeley_conservative,
annotator_ideology_conservative, annotator_ideology_slightly_conservative, annotator_ideology_neutral,
annotator_ideology_slightly_liberal , annotator_ideology_liberal, annotator_ideology_extremeley_liberal
annotator_ideology_no_opinion, annotator_race_asian, annotator_race_black, annotator_race_latinx,
annotator_race_middle_eastern, annotator_race_native_american, annotator_race_pacific_islander,
annotator_race_white, annotator_race_other, annotator_religion_atheist, annotator_religion_buddhist,
annotator_religion_christian , annotator_religion_hindu , annotator_religion_jewish,
annotator_religion_mormon, annotator_religion_muslim, annotator_religion_nothing, annotator_religion_oth
annotator_sexuality_bisexual, annotator_sexuality_gay, annotator_sexuality_straight, annotator_sexuality
```

Source code

```
library(knitr)
library(tidyverse)
library(broom)
library(htmltools)
library("tidyverse")
library("ggalluvial")
library("gridExtra")
# opts_chunk$set(echo=FALSE) # hide source code in the document


path <- setwd("..")
df <- read.csv(
  "data\\measuring_hate_speech.csv")
df_wo_red <- df
df_wo_red['annotator_race'] <- 0

df_wo_red <- within(df_wo_red, annotator_race[annotator_race_asian == 'True']<-'asian')
df_wo_red <- within(df_wo_red, annotator_race[annotator_race_black == 'True']<-'black')
df_wo_red <- within(df_wo_red, annotator_race[annotator_race_latinx == 'True']<-'latinx')
```

```
df_wo_red <- within(df_wo_red, annotator_race[annotator_race_middle_eastern == 'True']<-'middle_eastern
df_wo_red <- within(df_wo_red, annotator_race[annotator_race_native_american == 'True']<-'native_americ
df_wo_red <- within(df_wo_red, annotator_race[annotator_race_pacific_islander == 'True']<-'pacific_isla
df_wo_red <- within(df_wo_red, annotator_race[annotator_race_white == 'True']<-'white')
df_wo_red <- within(df_wo_red, annotator_race[annotator_race_other == 'True']<-'other')
# df_wo_red <- filter(df_wo_red, annotator_race != 0)



df_wo_red['target_race'] <- 0

df_wo_red <- within(df_wo_red, target_race[target_race_asian == 'True']<-'asian')
df_wo_red <- within(df_wo_red, target_race[target_race_black == 'True']<-'black')
df_wo_red <- within(df_wo_red, target_race[target_race_latinx == 'True']<-'latinx')
df_wo_red <- within(df_wo_red, target_race[target_race_middle_eastern == 'True']<-'middle_eastern')
df_wo_red <- within(df_wo_red, target_race[target_race_native_american == 'True']<-'native_american')
df_wo_red <- within(df_wo_red, target_race[target_race_pacific_islander == 'True']<-'pacific_islander')
df_wo_red <- within(df_wo_red, target_race[target_race_white == 'True']<-'white')
df_wo_red <- within(df_wo_red, target_race[target_race_other == 'True']<-'other')
# df_wo_red <- filter(df_wo_red, target_race != 0)

anno_wo_red<-df_wo_red


df_wo_red['annotator_religion'] <- 0

df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_atheist == 'True']<-'atheist')
df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_buddhist == 'True']<-'buddhist')
df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_christian == 'True']<-'christian')
df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_hindu == 'True']<-'hindu')
df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_jewish == 'True']<-'jewish')
df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_mormon == 'True']<-'mormon')
df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_muslim == 'True']<-'muslim')
df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_nothing == 'True']<-'nothing')
df_wo_red <- within(df_wo_red, annotator_religion[annotator_religion_other == 'True']<-'other')
# df_wo_red <- filter(df_wo_red, annotator_religion!=0)


df_wo_red['target_religion'] <- 0

df_wo_red <- within(df_wo_red, target_religion[target_religion_atheist == 'True']<-'atheist')
df_wo_red <- within(df_wo_red, target_religion[target_religion_buddhist == 'True']<-'uddhist')
df_wo_red <- within(df_wo_red, target_religion[target_religion_hindu == 'True']<-'hindu')
df_wo_red <- within(df_wo_red, target_religion[target_religion_jewish == 'True']<-'jewish')
df_wo_red <- within(df_wo_red, target_religion[target_religion_mormon == 'True']<-'mormon')
df_wo_red <- within(df_wo_red, target_religion[target_religion_muslim == 'True']<-'muslim')
df_wo_red <- within(df_wo_red, target_religion[target_religion_other == 'True']<-'other')

# df_wo_red <- filter(df_wo_red, target_religion != 0)



df_wo_red['annotator_sexuality'] <- 0
```

```
df_wo_red <- within(df_wo_red, annotator_sexuality[annotator_sexuality_bisexual == 'True']<-'bisexual')
df_wo_red <- within(df_wo_red, annotator_sexuality[annotator_sexuality_gay == 'True']<-'gay')
df_wo_red <- within(df_wo_red, annotator_sexuality[annotator_sexuality_straight == 'True']<-'straight')
df_wo_red <- within(df_wo_red, annotator_sexuality[annotator_sexuality_other == 'True']<-'other')

# df_wo_red <- filter(df_wo_red, annotator_sexuality!=0)


df_wo_red['target_sexuality'] <- 0

df_wo_red <- within(df_wo_red, target_sexuality[target_sexuality_bisexual == 'True']<-'bisexual')
df_wo_red <- within(df_wo_red, target_sexuality[target_sexuality_gay == 'True']<-'gay')
df_wo_red <- within(df_wo_red, target_sexuality[target_sexuality_lesbian == 'True']<-'gay')
df_wo_red <- within(df_wo_red, target_sexuality[target_sexuality_straight == 'True']<-'straight')
df_wo_red <- within(df_wo_red, target_sexuality[target_sexuality_other == 'True']<-'other')

# df_wo_red <- filter(df_wo_red, target_sexuality != 0)
# Annotator

 #Race
df_wo_red <- subset(df_wo_red, select = -c(annotator_race_asian, annotator_race_black, annotator_race_la

 #Religion
df_wo_red <- subset(df_wo_red, select = -c(annotator_religion_atheist, annotator_religion_buddhist, anno

 #Sexuality
df_wo_red <- subset(df_wo_red, select = -c(annotator_sexuality_bisexual, annotator_sexuality_gay, annota

#Target
df_wo_red <- subset(df_wo_red, select = -c( target_race_asian, target_race_black,
target_race_latinx, target_race_middle_eastern, target_race_native_american,
target_race_pacific_islander, target_race_white, target_race_other, target_religion_atheist, target_reli
target_religion_christian, target_religion_hindu, target_religion_jewish,
target_religion_mormon, target_religion_muslim, target_religion_other,
target_religion, target_origin_immigrant, target_origin_migrant_worker,
target_origin_specific_country, target_origin_undocumented, target_origin_other,
target_origin, target_gender_men, target_gender_non_binary,
target_gender_transgender_men, target_gender_transgender_unspecified, target_gender_transgender_women,
target_gender_women, target_gender_other,
target_sexuality_bisexual, target_sexuality_gay, target_sexuality_lesbian,
target_sexuality_straight, target_sexuality_other,
target_age_children, target_age_teenagers, target_age_young_adults,
target_age_middle_aged, target_age_seniors, target_age_other,
target_age, target_disability_physical, target_disability_cognitive,
target_disability_neurological, target_disability_visually_impaired, target_disability_hearing_impaired
target_disability_unspecific, target_disability_other, target_disability))

# Post
df_wo_red <- subset(df_wo_red, select = -c(text, platform, status, text))

# Redundant
df_wo_red <- subset(df_wo_red, select=-c(annotator_income_.10k,annotator_income_10k.50k, annotator_incom
annotator_income_.200k, annotator_ideology_extremeley_conservative,annotator_ideology_conservative,
```

```
annotator_ideology_slightly_conservative, annotator_ideology_neutral,annotator_ideology_slightly_liberal
annotator_ideology_liberal, annotator_ideology_extremeley_liberal,annotator_ideology_no_opinion, annotat
annotator_education_high_school_grad, annotator_education_some_college, annotator_education_college_grad
annotator_education_professional_degree, annotator_education_masters, annotator_education_phd, annotator
race_filter <- df_wo_red %>% filter(df_wo_red['target_race'] != 0)

p_target <- ggplot(race_filter, aes(x=target_race))+
  geom_bar(fill = 'dark blue', alpha =0.3) +
  labs(
    x = "Tragets Race",
    y = "Counter",
    title = "Number of Targets for each Race",
    subtitle ="Dark blue fill for targets",
  )+
  theme(axis.text.x = element_text(angle = 75,vjust = 0.5),
        axis.line = element_line( colour = "black")
  )

race_filter <- race_filter %>% filter(race_filter['annotator_race'] != 0)

p_anno <- ggplot(race_filter, aes(x=annotator_race))+
  geom_bar(fill = 'dark red', alpha =0.3) +
  labs(
    x = "annotator Race",
    y = "Counter",
    title = "Number of annotators for each Race",
    subtitle ="Dark red fill for annotator",
  )+
  theme(axis.text.x = element_text(angle = 75,vjust = 0.5),
        axis.line = element_line( colour = "black")
  )

grid.arrange(p_anno,p_target, ncol=2)
#("red", "blue", "yellow",'green')

race_filter <- within(race_filter, annotator_race[annotator_race == 'middle_eastern']<-'other')
race_filter <- within(race_filter, annotator_race[annotator_race == 'pacific_islander']<-'other')
race_filter <- within(race_filter, annotator_race[annotator_race == 'native_american']<-'other')
race_filter <- within(race_filter, annotator_race[annotator_race == '0']<-'other')
count_races <- count(race_filter, target_race, annotator_race)


ggplot(count_races,
       aes( axis1 = annotator_race, axis2 = target_race, y=n)) +
    geom_alluvium(aes(fill = target_race, color = target_race),
                  width = 4/12, alpha = 0.3, knot.pos = 0.4) +
    geom_stratum(width = 1/4,color = "grey") +
    scale_fill_manual(values=c("red", "blue", "yellow", "purple",
                               "cadetblue4", "#95CC5E", "lightgrey", "green"))+
    scale_color_manual(values=c("red", "blue", "yellow", "purple",
                                "cadetblue4", "#95CC5E", "lightgrey", "green"))+
    geom_text(stat = "stratum", aes(label = after_stat(stratum)), size=2.53) +
    scale_x_continuous(breaks = 1:2, labels = c("Annotators", "Target"))  +
```

```
    theme_minimal() +
    theme(
        legend.position = "left",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text.y = element_blank(),
        axis.text.x = element_text(size = 11, face = "bold"),
        )+
    labs(
        title = "Distribution of annotators race vs target race",
      subtitle = "Emphasizes the Annotators Race and the race of the target they tagged",
      y = 'Races'

    )
p1 <- ggplot(df_wo_red, aes(x=annotator_age))+
  geom_density(fill = 'dark red', alpha =0.3) +
  labs(
    x = "Age",
    y = "Density",

    title = "Density of annotators age",

  )+
  theme(axis.text.x = element_text(angle = 75,vjust = 0.5),
        axis.line = element_line( colour = "black")
  )
p2 <-  ggplot(df_wo_red, aes(x=annotator_age))+
  geom_bar(fill = 'dark red', alpha =0.3) +
  labs(
    x = "Age ",
    y = "Count",

    title = "Number of annotators age",

  )+
  theme(axis.text.x = element_text(angle = 75,vjust = 0.5),
        axis.line = element_line( colour = "black")
  )

grid.arrange(p1,p2, ncol=2)
ggplot(race_filter, aes (x = hate_speech_score, fill = annotator_race))+
  geom_density(alpha =0.2)+
  labs(
    title = "Density for hate speech score by race",
    subtitle = "color is for the annotator race",
    x= "hate speech score",
    y = "density",
    fill = "Annotator Race")+
  theme_minimal()
white_anno <- race_filter %>% filter(annotator_race == 'white'&
                                       target_race != "other"&
                                       target_race != "pacific_islander"&
                                       target_race != "native_american"&
```

```
                                        target_race != "middle_eastern")


black_anno <- race_filter %>% filter(annotator_race == 'black'&
                                        target_race != "other"&
                                        target_race != "pacific_islander"&
                                        target_race != "native_american"&
                                        target_race != "middle_eastern")

asian_anno <- race_filter %>% filter(annotator_race == 'asian'&
                                        target_race != "other"&
                                        target_race != "pacific_islander"&
                                        target_race != "native_american"&
                                        target_race != "middle_eastern")

latin_anno <- race_filter %>% filter(annotator_race == 'latinx'&
                                        target_race != "other"&
                                        target_race != "pacific_islander"&
                                        target_race != "native_american"&
                                        target_race != "middle_eastern")


plot_w <- ggplot(white_anno, aes (x = hate_speech_score, fill = target_race))+
  geom_density(alpha =0.2)+
   scale_fill_manual(values=c("red", "blue", "yellow",'green'))+
  labs(
    title = "White annotator",
    subtitle = "Density of score by target race",
    x= "Hate Speech Score",
    y = "Density",
    fill = "Traget")+
  theme_minimal()


plot_b <- ggplot(black_anno, aes (x = hate_speech_score, fill = target_race))+
  geom_density(alpha =0.2)+
   scale_fill_manual(values=c("red", "blue", "yellow",'green'))+
  labs(
    title = "Black annotator",
    subtitle = "Density of score by target race",
    x= "Hate Speech Score",
    y = "Density",
    fill = "Traget")+
  theme_minimal()

plot_a <- ggplot(asian_anno, aes (x = hate_speech_score, fill = target_race))+
  geom_density(alpha =0.2)+
   scale_fill_manual(values=c("red", "blue", "yellow",'green'))+
  labs(
    title = "Asian annotator",
    subtitle = "Density of score by target race",
    x= "Hate Speech Score",
    y = "Density",
```

```
    fill = "Target")+
  theme_minimal()

plot_l <- ggplot(latin_anno, aes (x = hate_speech_score, fill = target_race))+
  geom_density(alpha =0.2)+
   scale_fill_manual(values=c("red", "blue", "yellow",'green'))+
  labs(
    title = "Latin annotator",
    subtitle = "Density of score by target race",
    x= "Hate Speech Score",
    y = "Density",
    fill = "Target")+
  theme_minimal()

grid.arrange(plot_w, plot_b, plot_a, plot_l, ncol=2)

cat(readLines('../data/README.md'), sep = '\n')
```