Analytics for Social Good:
Addressing Social Biases in Algorithmic Systems
Lab Session
Trainer: Jahna Otterbacher

Objective: In this exercise, we shall examine how four popular computer vision algorithms – in particular, content tagging APIs – interpret images depicting people. In particular, we shall examine their output tags for evidence of systematic gender- and race-basedl biases.

Method: We shall propose a "digital controlled experiment," by submitting a set of people images to each algorithm, and analyzing the output (i.e., word tags) through textual and then statistical analysis.

Dataset: The Chicago Face Database (CFD)[1] is a free resource consisting of 597 high-resolution, standardized images of diverse individuals between the ages of 18 and 40 years. It is designed to facilitate research on a broad range of psychological phenomena, (e.g., stereotyping and prejudice, interpersonal attraction). Therefore, it provides extensive data about the depicted individuals. The database includes both subjective norming data (i.e., ratings for perceived attributes, reported on a scale from 1 to 7[2]), and objective physical measurements (e.g., nose length/width), on the pictures[3]. For our purposes, a significant benefit is that the individuals are depicted in a similar, neutral manner; if we were to evaluate images of people collected "in the wild" we would have images from a variety of contexts with varying qualities. In other words, using the CFD enables us to study the behavior of the tagging algorithms in a more controlled manner.

Materials: For the purposes of the 90-minute session, the output from the four taggers has already been obtained for you. You can find the full dataset, the Python wrappers used to process the CFD images, as well as the R resources here https://tinyurl.com/ChipSetSocialGood.

(If you want to download the dataset, please see the "ARCHIVED" directory.)

Tools: For our analysis, we'll be using **R Studio**, as well as the following R libraries:

Text manipulation

- tidyr useful functions for "tidying up" messy data
- dplyr more functions for manipulating data
- stringr functions specificly for handling strings (e.g., regular expressions)

Statistical analysis

- lsr contains some useful functions for statistical analyses and computing effect sizes
- mass contains many basic statistics functions

[1] D. S. Ma, J. Correll and B. Wittenbrink, 2015. The Chicago face database: A free stimulus set of faces and norming data. Behavior Research Methods, 47(4), 1122-1135.
[2] The CFD provides the mean scores given by over 30 raters for each photograph.
[3] Physical measurements are reported in pixels (i.e., are measured from photos.)

Exercise 0: Interpreting gender

While the tagging algorithms we are testing are not specifically designed for gender recognition, it is easy to see to many of the output tags – across all four taggers – imply the depicted person's gender (e.g., man, boy, girl, lady). Therefore, one interesting thing to consider if how often the taggers correctly use masculine and feminine tags, and whether or not this is correlated to the depicted person's race and gender.

1a) For each API, produce the confusion matrix showing the *inferred gender* versus the *true gender* of depicted persons. Let's assume that when an API uses tags of only one gender, that it implies the person's gender (i.e., only tags such as "man" and "boy" are used to describe a given image, and no feminine tags).

1b) Conduct an appropriate statistical analysis to examine whether the taggers' accuracy on gender inference is correlated to the (actual) race and gender of the depicted persons.

Step 0: Load the libraries we need, explore the dataset and perform data cleaning as needed.
→PrepareData.R

Step 1: Since we know nothing about the taggers' behaviors, we'll start by discovering the set of all possible tags (i.e, the tag vocabulary). The following will produce the file "unique_tag.csv" in the working directory.
→GetTagLexicon.R

Step 2: Now we can have a look at the tag vocabulary, in order to create a list of feminine and masculine reference words. This is not entirely an objective process! For instance, we might argue for or against the inclusion of words such as "necktie" or "beard" in the list of masculine words, whereas other words such as "man" or "boy" are much less contentious. A best practice here is to ask a number of independent judges to analyze the words for you.

Step 3: The next step is to match each wordlist against the set of tags produced by each tagger for each image (i.e., the variables "Clarifai" "Microsoft" "Watson" and "Imagga" in the original dataframe). We'll create a binary variable for each tagger, indicating whether or not the tagger has correctly inferred the gender of the depicted individual. Then we'll produce the confusion matrix.
→ Match-feminine-masculine.R

Step 4: Finally, let's conduct a Test of Independence to see if the depicted person's race is correlated to the taggers' abiity to correctly infer gender. Given that we've seen that the taggers' associate nearly every individual with masculine tags, and that the data set is balanced in terms of both gender and race, we don't expect any significant results – but let's double check it.
→ Stats-ToI.R

Exercise 1: Who is physically attractive according to the taggers?

Step 1: Now let's revisit the vocabulary of tags and create a list of words that refer to a depicted person's physical attractiveness.
→ `attractive_words.txt`

Step 2: Once we have our list of words, we'll need to see which set of tags contains them. We'll do something really simple here – grep for each word. We can then compute a score that reflects the extent to which a given tagger uses these words when describing an image, e.g., the proportion of total words used by the tagger that are in our list of "attractive" words.
→ `MatchAttractive.R`

Step 3: Once we have our scores, we can first examine the extent to which the taggers' use of attractive words correlates to the scores assigned by human raters (i.e., the "Attractive" score from the Chicago Face Database). Finally, we can conduct some simply statistical analyses (ANOVA) to examine for evidence of gender- and race-based bias. In other words, we can see whether the use of attractiveness words is correlated to the race/gender of the depicted person.
→ `Stats.R (commands only)`
→ `Stats-with-Output.R`