National University of Sciences and Technology (NUST)
School of Electrical Engineering and Computer Science

Machine Learning

Semester Project

Initial Report

Dataset Bias-Skin Color Classifier

Group Members.
Sannan Abbasi
Junaid Ali
Amur Saqib Pal

Abstract.

While developing an ML model, there are different types of biases that can exist in a dataset. The problem with having a biased dataset is that the model ends up acting with biases as well. This can cause problems such as skewed results, results with low accuracy and analytical errors. In this project, we are exploring sample and racial data biases that exist in different search engines (Google, Bing, DuckDuckGo) when we make queries regarding different nationalities. Furthermore, we will also attempt at developing a skin colour classifier which would accurately predict the nationality of a person through the colour of that person's skin.

Progress.

Collection.

For the first part of our project in which we are to explore data bias in search engines, we had to create a data set of the images that are available on different search engines when we searched for a particular nationality. The three search engines, as mentioned above, are Google, Bing and DuckDuckGo.

The 10 countries we chose to gather images for are.

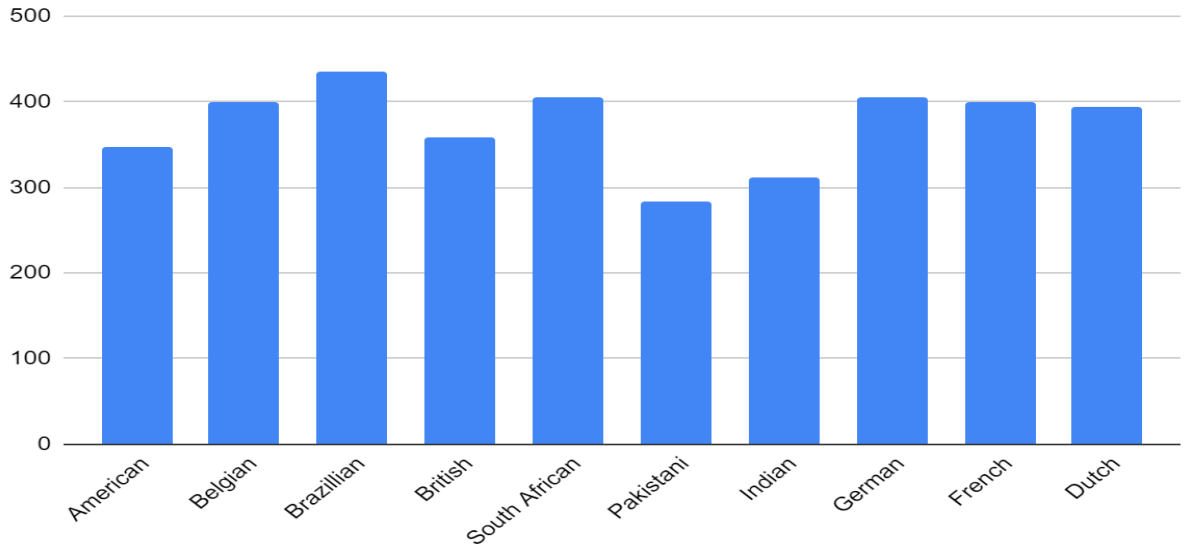
1. America
2. England
3. Brazil
4. South Africa
5. Netherlands
6. Belgium
7. France
8. Germany
9. Pakistan
10. India

The reason we chose these countries was so we could have a data set that had equal representation of the three skin colours that we were going to study i.e Black, Brown and White.

The images were downloaded using an extension available on Google Chrome called “Download All Images”. We wanted to get at least 1000 images per nationality from each search engine. However, the search engines had a limited amount of images to display for each nationality. For example, we only got on average 373 images from Bing for each nationality. The number of images we downloaded from each search engine is given below.

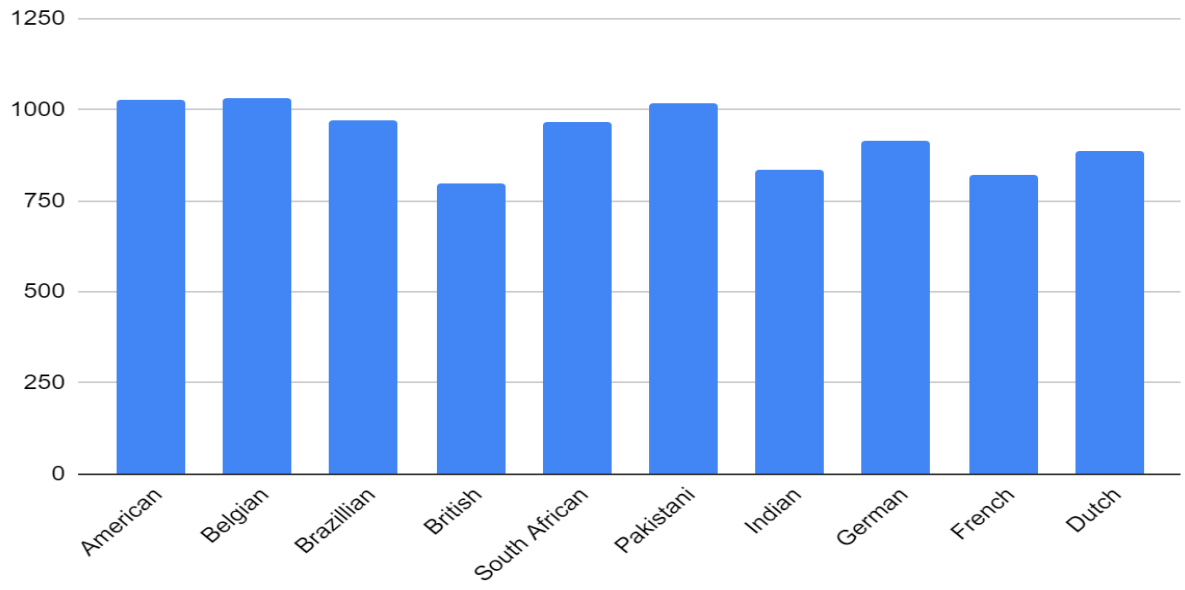
Bing.

Total Images: 3739



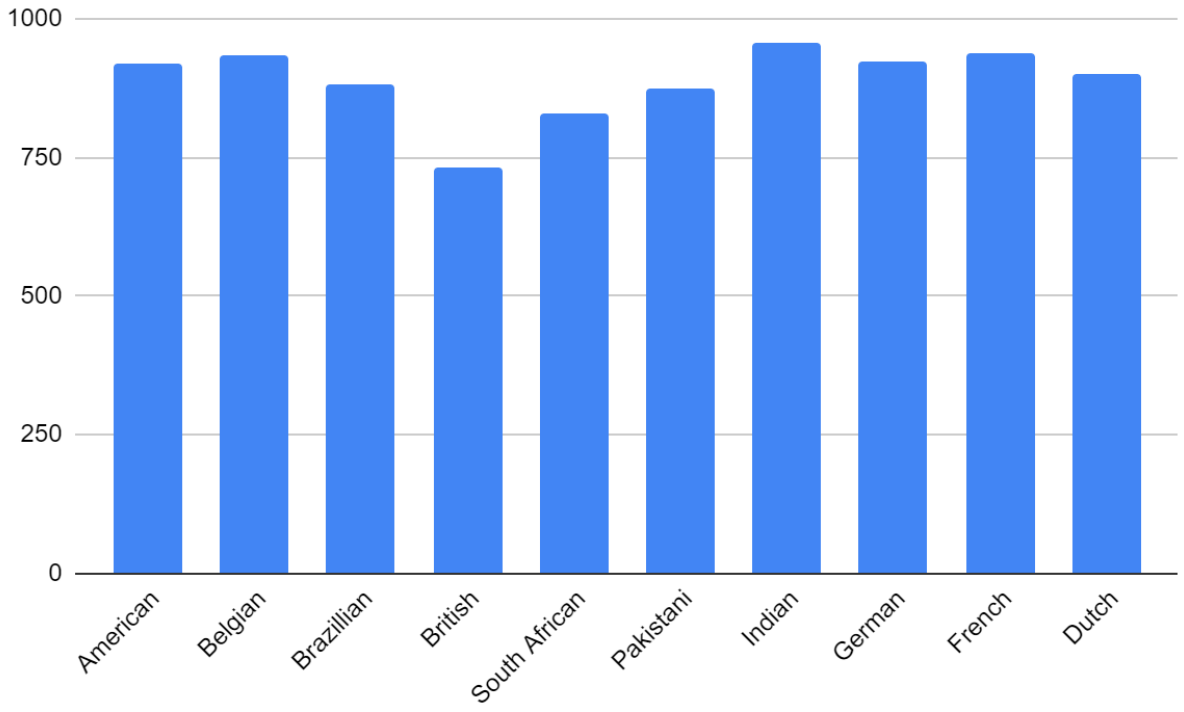
Google.

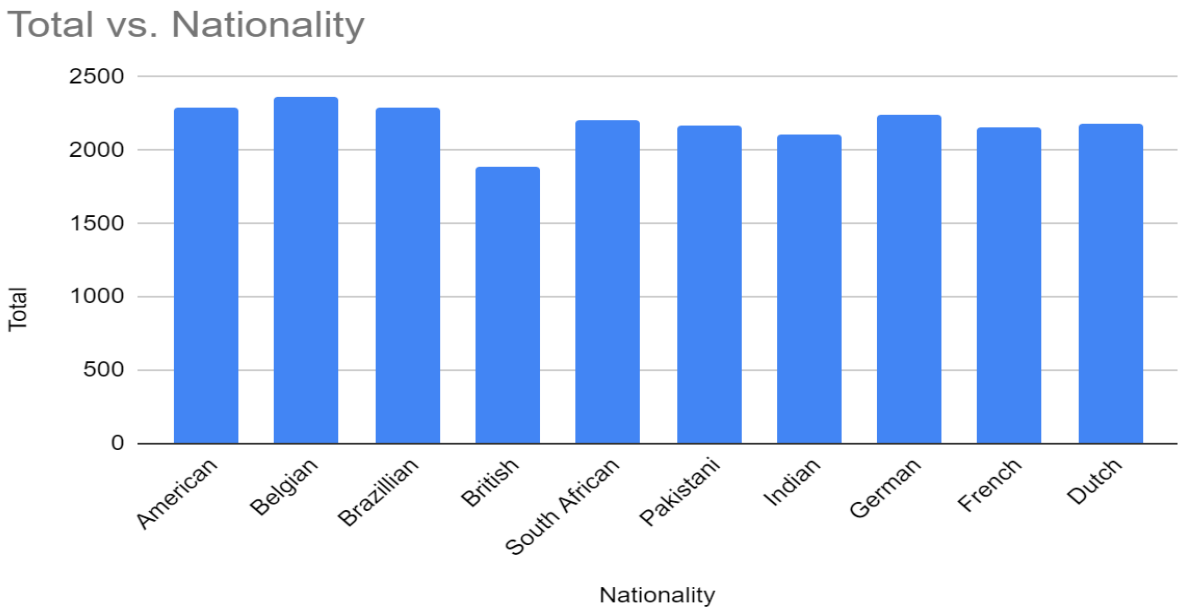
Total Images: 9273



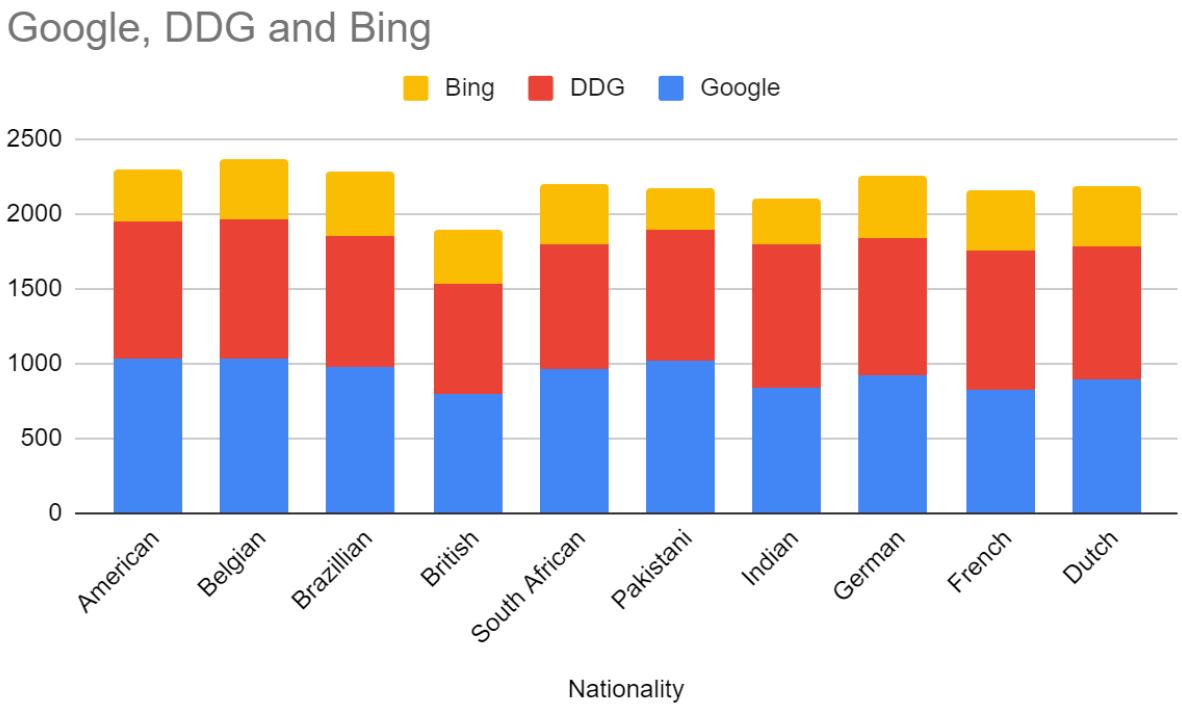
DuckDuckGo.

Total Images: 8889





The Net Total: 21901



Cleaning and Labelling the Data.

The next step was to clean the data set we had downloaded. Till now we have done 5 nationalities.

1. German
2. Brazillian
3. French
4. Belgian
5. South African

In the cleaning process, we got rid of all the images that had little to no people. We also got rid of the images that were not clear enough to label them with any skin colour.

Next, we labelled the images that were left and divided them in three skin colours i.e Black, Brown and White. This has been done for the above-mentioned nationalities.

This is the google drive link where the data set is stored along with a spreadsheet containing the information on cleaning the data.

https://drive.google.com/drive/folders/1QD0JXODPt-gbeg-BJH6OAhrtb7-cPBZ8?usp=sharing

Next Steps?

We will continue cleaning and labelling our data for the remaining nationalities. Moving further we will work on our skin colour classifier. Our idea is to use a KNN classifier to develop our predictive model.