**Capstone Research Project Proposal**

This project studies how training language models on a variety of languages affects racial and gender bias in in these models. Specifically, I ask how the choice of languages used in training different versions of a model impacts racial and gender bias in natural language processing tasks? To answer this I examine the shifts in racial and gender bias across BERT models trained for a variety of languages.

While language models, such as BERT, have revolutionized NLP tasks, there are concerns that these models are trained to exhibit bias unintentionally. In a study done by Caliskan et al. (2017), the authors evaluated the cosine similarity between a subset of words pertaining to a variety of categories. Their tests discovered the model’s racial and gender biases. Through this study, the authors showed not only that the GloVe model exhibits these commonly held biases, but also that these biases could lead to disproportionate representation in automated decision-making processes.

I propose a similar design to Caliskan et al. in which I evaluate the cosine similarity of a subset of target words and concepts to determine if the selected models exhibit bias. All the models evaluated will be trained in both English and in at least one other language. This ensures that when running the tests for cosine similarity, no translation will be needed. The models being evaluated are available from Hugging Face. The data I will evaluate these models on will change depending on the kind of bias I want to look at.

To measure bias in these models, I test how similar a set of words is to another. For example, in the test for gender bias, I evaluate the cosine similarity between two sets of the most popular names for both men and women respectively and two sets of job labels, one consisting of STEM-related jobs and one of non-STEM-related fields. In doing so, I compare how closely related these sets of words are to each other, and whether one set of names is more closely linked to one of the two sets of job labels. (having an issue trying to put thoughts into words, look to comments for more details)

To test for biases, I will create a set of both target words and a set of concepts in my evaluations of cosine similarity. The set of target words used in each test will be consistent throughout to keep results consistent. These two sets will consist of words which are considered Pleasant (e.g. happy, kind, enjoyable) and Unpleasant (e.g. sad, angry, disgusting). I will first test for race using subsets of names that are commonly classified as European-American, African-American, Asian-American, and Latin-American, and evaluating the cosine similarity of these subsets with both the group of Pleasant and Unpleasant words. I also evaluate individual generic terms, such as Latina/Latino, African, etc. with the same group of target words. I will then test for gender bias following a similar format, in which I will gather two sets of the most common names in the US for both men and women respectively, and evaluate the cosine similarity between each set and the set of target words. With regards to gender bias, I will also evaluate the cosine similarity between these sets of names and two sets of job labels. These two sets of labels will consist of labels pertaining to STEM careers, such as engineer or data scientist, and non-STEM fields, such as journalist or teacher.

Sources

Caliskan, Aylin, et al. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” *Papers With Code*, 25 Aug. 2016, paperswithcode.com/paper/semantics-derived-automatically-from-language.

“Bert-Base-Uncased · Hugging Face.” *Bert-Base-Uncased · Hugging Face*, huggingface.co/bert-base-uncased. Accessed 17 May 2023.