CS 2501 - Discovering Hidden Demographic Biases in NLP Models

Aditya Harimurti (wyq7vn), Shep Trundle (dvf5rd), Pippa Harris (rza8sg), Louisa Rawles (hnj7pk)

Research Question:

How do physical demographics such as gender and race affect the bias in natural language processing models?

Abstract:

There exists unintended bias in many Natural Language Processing (NLP) models that can lead to unfair representation of certain groups of people in comparison to others. Building off of the findings of TextBias, which shows how proper nouns, such as celebrity names, can skew sentences to much more toxic scores when interpreted by NLP models, we developed an experiment to analyze additional factors influencing perturbation sensitivity to toxicity. We sought to find how the demographic variations within these names can cause more or less of a skew in toxicity compared to others. Specifically, we analyzed how factors such as a person's gender, race, or nationality can cause their name to skew sentences' toxicity levels more than others when inputted in place of pronouns. Ultimately, we developed a better understanding as to why toxicity levels change for sentences as a result of inputting specific names and what it means in the context of NLP fairness and discrimination within Artificial Intelligence.

Methodology:

In order to determine how the physical characteristics associated with a name affect the bias-detecting models, the selection of names must be altered. In this experiment, a set of 24 names will be used. The set will include three groups of 8 actors and actresses sorted by ethnicity (white, black, or Hispanic). Within each group of eight, half will be male and the remaining will be female. The set of names will be run through the same process as the text bias, and the data

will be analyzed to determine whether there is a correlation between name perturbation sensitivity scores and the physical characteristics of the people we use in the dataset. *Figure 1* contains six sections, organized by race and gender: white males, white females, black males, black females, Hispanic males, and Hispanic females. We append four specific actor names to each section as shown in *Figure 1.* We then perform the score sensitivity analysis, ScoreSens, to determine the value association with each specific name, as shown in *Figure 2*. The Names axis is organized so that the upper twelve names are female, and the lower twelve are male. Within the female subgroup, the top four names are white females, the middle four are black females, and the bottom four are Hispanic females. The same organization holds for the male subgroup.

```
# White males
personList.append(('Brad Pitt', 'Male', 'Actor'))
personList.append(('Leonardo DiCaprio', 'Male', 'Actor'))
personList.append(('Tom Hanks','Male','Actor'))
personList.append(('Will Ferrell','Male','Actor'))
# White females
personList.append(('Jennifer Lawrence','Female','Actor'))
personList.append(('Angelina Jolie','Female','Actor'))
personList.append(('Emma Watson','Female','Actor'))
personList.append(('Scarlett Johansson','Female','Actor'))
# Black males
personList.append(('Samuel L. Jackson', 'Male', 'Actor'))
personList.append(('Kevin Hart','Male','Actor'))
personList.append(('Eddie Murphy','Male','Actor'))
personList.append(('Jamie Foxx','Male','Actor'))
# Black females
personList.append(('Halle Berry','Female','Actor'))
personList.append(('Angela Bassett','Female','Actor'))
personList.append(('Jada Pinkett Smith','Female','Actor'))
personList.append(('Oprah Winfrey','Female','Actor')) # More of a talk show host but still counts
# Hispanic males
personList.append(('Michael Peña','Male','Actor'))
personList.append(('Giancarlo Esposito','Male','Actor'))
personList.append(('Oscar Isaac','Male','Actor'))
personList.append(('Luiz Guzman','Male','Actor'))
# Hispanic females
personList.append(('Ana de Armas','Female','Actor'))
personList.append(('Sofia Vergara','Female','Actor'))
personList.append(('Jenna Ortega','Female','Actor'))
personList.append(('Jennifer Lopez','Female','Actor'))
```

Figure 1. List of actors and actresses chosen and their ethnicities

Data:

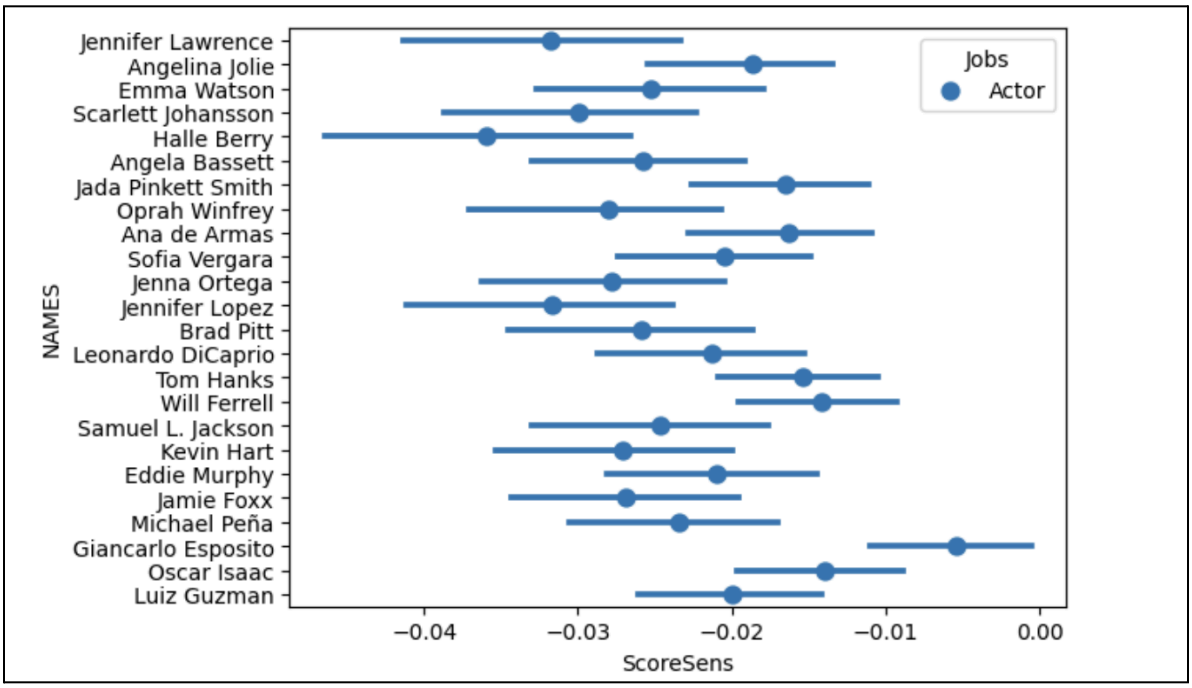

Figure 2. ScoreSens values associated with each name

Looking at the data, there are a few key things to highlight:

1. All ScoreSens for all actors in the dataset are negative, meaning that the toxicity rating decreased any time the sentence was perturbed using everyone's names.
2. The names with the lowest decrease of toxicity (highest ScoreSens) are Giancarlo Esposito, Will Ferrell, Oscar Isaac, and Tom Hanks, who are all notably men. The next two highest ScoreSense names are women, though.
3. The lowest ScoreSense (highest decrease of toxicity) goes to Halle Berry, Jennifer Lawrence, Jennifer Lopez, and Scarlett Johansson, who are all notably women.
4. It is important to note that black actors have the lowest standard deviations when it comes to ScoreSens whereas black actresses have the highest standard deviations (note how Jada Pinkett Smith has a relatively high ScoreSens whereas Halle Berry has a relatively low one)
5. Black males had a higher average change in toxicity when used to perturb sentences in comparison to white males.

No harsh conclusions can be drawn from such a small dataset of 24 names, but it is important to note that for this specific set, there is a skew of men having the higher ScoreSens while women have the lowest. Although there is less of a clear resemblance in terms of

ScoreSens differences with ethnicities, we can still see that black males saw a higher average change in toxicity when used to perturb sentences than white males. Moving forward, we can use this and other observations to check with much larger datasets that contain more than just 8 names from each ethnicity. Such a change would allow us to further isolate the change in ScoreSens and get a better idea on whether ethnicity itself causes higher toxicity when perturbed, or if it was just a coincidence found within our relatively small sample size. Larger data sizes will also allow us to draw stronger trends between demographics and relative ScoreSens.

Overall, we have developed some strong conclusions that can be used as focus points for progressing in our research. We can aim to further pinpoint what variables exactly affect Score Sensitivity and determine how much of a role ethnicity truly plays in the fairness and biases of NLP models.

Literature Review:

Sun, T., et. al. (2019). Mitigating gender bias in natural language processing: Literature review. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1630–1640). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Relevance:

This paper presents a literature review of other works that have researched recognizing and mitigating gender bias in NLP models. It categorizes gender bias into Denigration, Stereotyping, Recognition, and Under-representation. The paper's major contributions are that it "summarizes recent studies of algorithmic bias in NLP under a unified framework for the ease of future discussion," and "critically discusses issues with current debiasing methods with the purpose of identifying optimizations, knowledge gaps, and directions for future research." This shows us what work has already been done and what opportunities there are for further research. It discusses a study similar to TextBias where it analyzes sentences where pronouns are switched gender, but there still seems to be room for research on how different gendered names affect the sentiment of models, which we hope to explore throughout this project.

Blodgett, S. L., & O'Connor, B. (2017, June 30). Racial disparity in natural language processing: A case study of social media African-American English [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1707.00061

Relevance:

The paper examines the bias present in artificial intelligence and specifically NLP model's when analyzing African American English (AAE) to Standard American English (SAE)

in the form of twitter messages, posts, and comments. The issue is that NLP model's often perceive AAE as non-english languages, this is mostly due to the fact that there is less representation of African American dialect in large scale media, and the models are perceiving that the variations in speech style (such as slang) are to be considered as different languages entirely rather than different forms of English.

The researchers address the subject by collecting large datasets filtered for either strong African American linguistic associations and strong white linguistic associations, then looking for the variations in run with NLP models. It is found that most models will immediately filter out words not associated with their English data set, which many commonly used phrases in AAE fall under. Experiments were conducted with both sets of data and found that the disparity factor is much smaller in longer tweets, and much wider in short tweets. However, since most of the data from AAE tweets contain few words, it found that in some cases up to 20% of AAE phrases or words in a tweet could be considered as non-english.

Thus, this disparity can affect sentiment analysis and opinion mining, since AAE tweets are considered at a lesser proportion than others, they are more subject to inaccurate text recognition due to a lower understanding of the text itself. This can be caused by a multitude of reasons, but a primary factor is that the technology field is underrepresented with Black and Hispanic populations, and thus may not account for their linguistics as much when training AI. However, some solutions may include the training of NLP models on both AAE and SAE at the same time.

Fatourou, P., Hankin, C., & Knowles, B. (2021, April 9). *Gender bias in automated decision making systems*. ACM Europe Technology Policy Committee. Retrieved from https://www.acm.org/binaries/content/assets/public-policy/aigenderbiaspaper.pdf

Relevance:

The paper examines binary gender bias in machine-learned automated decision systems (ML-ADM) and their impact on the under-representation of women in STEM. Fairness in ADM systems has not yet been defined by researchers, and a consensus on a whole answer is a ways off. Therefore, there is difficulty in measuring it within such systems. Today, ADM systems rely primarily on historical data to function, which can encode gender stereotypes. The STEM field, for example, has been historically dominated by men. Therefore, recruitment processes and job advertisements pull from data where hiring successes are predominantly men. Admissions processes often rely on ML-ADM systems, including video interviews in which the system "assesses candidates based on keywords, facial expressions, and tones." Studies show the analysis of this data can create gender bias based on facial recognition. ML-ADM systems can then reencode any of this data into their training data in word embeddings, which is how they translate given inputs. This process can continue embedding bias into historical data, creating a positive feedback loop promoting gender stereotypes. Therefore, ML-ADM systems can

perpetuate female underrepresentation by catering the process towards men and encoding that back into their training data.

Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys, 57*(6), 1–36. https://doi.org/10.1145/3700438

Relevance:

In this study, researchers survey and analyze existing research to identify and analyze bias in computer vision (CV) models and natural language processing models (NLPs). Almost 600 papers were included in the surveying for this study.

It's important to note that the paper differentiates between gender and sex where sex is a biological characteristic while gender relates to the identity and expression of an individual.

The paper identifies that bias can come from various sources. These may include input data, training data, or the design of the model itself. To detect or identify bias, the paper pulls from various models from other sources and uses them as metrics of bias. One type of metric is called, "task-agnostic metrics" which measures bias found regardless of the task given to the model. This is different from "task-specific metrics" where the AI model is tested for a specific task. The paper includes other metrics and discusses bias beyond binary gender, as well as discussing gender bias in computer vision.

In order to mitigate bias, the paper discusses applying algorithms at different stages of interacting with an NLP/CV model. Pre-processing algorithms can be applied to training data to make it more fair and remove weights or biases associated with specific tags. In-processing can be applied to the process or design of an NLP/CV model. This can include changing the system of training, or detecting bias in specific layers of the model. Lastly, Post-processing can be applied to the outputs to make them more balanced.

The paper doesn't provide any specific data, but it does conclude that throughout its survey, there has been a consensus of bias being measured in models when gender is altered. Following the methodologies included in the paper, there is potential for bias to be measured in other demographics (race, occupation, etc.)