CSE 246 Project Proposal

Evaluating Biases for GPT-3

Farid Zakaria <fmzakari@ucsc.edu>
JunYi Yu <jyu184@ucsc.edu>

Motivation

OpenAI is an artificial intelligence research laboratory that has released groundbreaking AI and APIs to access these tools to the general public that have attracted widespread attention. One of the AI APIs offered by the laboratory is GPT-3 which is a language model that produces text for a variety of scenarios: classification, summarizing, translation etc..

Due to the popularity of the API due to its ease of use, it's becoming found increasingly in many of the products that are in use today. For instance, Github has recently launched Copilot which suggests code completion that utilizes GPT-3, albeit a model trained on coder rather than non context-free languages (i.e. English). Given this popularity, it is important to understand the ethical implications of relying on AI models to produce text or if they are involved in decision making from a responsible data science perspective.

Proposal

Our group proposes a series of simple experiments targeted against GPT-3 to evaluate if the model has any biases; specifically we will target gender bias for the experiment. Our group will construct a series of job postings for roles that are traditionally associated with a specific gender, such as women in nursing, and ask the model to select between two job candidates. In both cases, the provided resume will be **identical** except for gender qualifiers and their name.

The names for the candidates will be selected from the US birth dataset which is offered as a public dataset accessible via BigQuery (bigquery-public-data.usa_names.usa_1910_2013). We will remove any names that exist in both genders to make the experiment simpler and remove any confusion for names that may exist amongst both genders in the US.

There has been some work in gender bias evaluation in GPT-3 [1][2], although in both cases the research looked at the generated text produced by GPT-3. We believe the experiment here to be novel in that it asks GPT-3 to select amongst the two candidates producing a simple binary result to evaluate.

Learning Outcomes

1. Learn more about AI/ML in general
2. Learn about OpenAI and the GPT-3 API and it's functionality
    a. Learn to access and use the API via one of their SDKs; likely Python.
3. Manually curate or produce a data pipeline to clean and curate the names used for the experiment.
    a. Learn more about the BigQuery public datasets offered and using the API
4. Evaluate OpenAI GPT-3 and discover if the trained model has any gender bias

Timeline

Preparation for Project
Due Date: Oct 14
Read related work literature like gpt-3 overview and Open API usage document.

Collect Resume & Job Postings
Due Date: Oct 25
Collect several resumes and related job postings. Collecting is not enough, we prefer to fabricate based on examples. After collection, we need to turn them into pure and clean texts.

Extract Enough Names from US birth dataset & Dataset Preparation
Due Date: Oct 28
Extract names from US birth dataset. The selection of names should target names with strong gender implication. The basic policy is we don't choose names that occur with both genders. After extraction, we will start to arrange names from different genders into the same resumes.

Mid Point Presentation
Due Date: Nov 3
At midpoint, our deliverable should be a complete dataset. And hopefully, we can start training the model to predict, then there will be a demo like, input texts, output a rating of resume.

Get Enough Accuracy of Resume Classification
Due Date: Nov 14
Try to make predictions based on gpt-3.

Detect possible gender bias
Due Date: Nov 21
Check the classification result, detect if certain gender will get a higher rating than another.

Stretch Goal: Any possible technical solutions to mitigate gender bias?

We are assuming that we will get gender bias from classification results. If that's true, since we have learned some technical solutions in the course, we can try to find a solution to mitigate the bias.

Final Presentation
Due Date: Dec 1
In the presentation, we will display the classification result and talk about the project process, and give a clear conclusion that if there is gender bias in resume screening.

Final Write-up
Due Date: Dec 5
Our deliverable would be several scripts and a summary paper of research results.