CS 7313 – Project Proposal (Track 2)

Authors: Andrew Scouten, Tanha Tahseen

Paper: https://aclanthology.org/2023.acl-long.84/
Code & Data: https://github.com/myracheng/markedpersonas

The *Marked Personas* paper investigates how Large Language Models (LLMs) like GPT-3.5 and GPT-4 produce stereotypes, especially those related to race and gender, when they are asked to describe people. Their goal was to measure these biases without any manually labeled data. To achieve this, the authors introduced *Marked Personas* - a prompt-based method to measure stereotypes in LLMs for intersectional demographic groups. The framework is two-stepped; First, the model is asked to generate short natural-language self-descriptions ("personas") for both marked groups (e.g., *Black woman*, *Asian man*) and unmarked or default groups. Next, the words are identified that appear significantly more often in the marked groups than in their unmarked counterparts, revealing stereotypical patterns in the model's language, using the Fightin' Words log-odds method supplemented by VADER sentiment analysis and a Support Vector Machine (SVM) classifier.

Problem Statement

The *Marked Personas* paper does not explore how these biases are encoded within the model's internal representations or whether alternative sentiment methods yield consistent results. This project extends *Marked Personas* by applying unsupervised clustering and feature visualization (e.g., t-SNE, UMAP, attention map visualization) to examine latent structure in persona embeddings and by testing additional sentiment techniques such as BERT and RoBERTa based sentiment classifiers. Together, these analyses aim to reveal how social bias manifests in the feature space of modern LLMs.

Methodology

1. Dataset and Embeddings

We will use the released *Marked Personas* dataset, which includes persona prompts and corresponding LLM-generated text. Sentence-level embeddings will be extracted for both prompts and generated text using pretrained transformer encoders (e.g., BERT or RoBERTa) to capture features associated with each persona.

2. Dimensionality Reduction and Feature Visualization

We plan to apply dimensionality reduction and clustering (t-SNE, UMAP, K-Means, DBSCAN, etc.) on the extracted features to help analyze the word embedding space, enabling visualization of how persona-related features organize within an LLM's latent representation. These plots could then be organized by demographic attributes and sentiment to assess clustering or separation indicative of bias. We can also better visualize an LLM's features by using interpretability tools (e.g. BERTViz, Captum, Ecco) to visualize attention distributions and compute token-level understandings for persona descriptors and generated text.

3. Sentiment Modeling and Word Importance

We will utilize similar models as to those used in feature extraction for our sentiment analysis, helping us gain an understanding of bias / intent in the model's response. By leveraging attention heatmaps, clustering, and other word importance tools, our analysis will identify which words or identity markers most influence representation geometry and sentiment outcomes.

4. Bias Quantification and Correlation

Finally, we correlate cluster structure, attention weights, and sentiment predictions to test whether the *Marked Personas* bias patterns are reflected in internal representations. Quantitative metrics (e.g., inter-group distance, sentiment variance, feature importance) and qualitative visualizations together will demonstrate how bias is encoded and expressed in modern LLM embeddings.

Potential Applications

Token Classification:

- https://huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll03-english

Similarity:

- https://huggingface.co/google/embeddinggemma-300m

Text Classification (Sentiment):

- https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
- https://huggingface.co/GroNLP/hateBERT
- https://huggingface.co/Falconsai/intent_classification
- https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Evaluation Plan

Our evaluation will have two components: replication validation and extension analysis. To verify the correctness of our replication, we will reproduce the key quantitative analyses from the *Marked Personas* paper and compare our numbers with the results reported in Tables and Figures of the paper (e.g., stereotype-word percentages, sentiment distributions, and significant-word lists). We will reproduce the SVM experiment and compare our classification accuracy with the original paper (reported 0.92–0.96 accuracy). After replicating the study, we can then apply our methods and compare the results with the paper's findings. Some of the paper is objective, where we can directly compare our metrics to theirs, however there are portions of the paper that will be inherently subjective depending on how we define "bias". Therefore we will measure our success as both the reproduction of the study and the implementation of our proposed methods along with the depth of our analysis.