Veronica Lin
CS 232 Artificial Intelligence Final Project
Prof. Carolyn Anderson
May 16, 2022

Investigating Cultural Assumptions in GPT-3 Generated Predictions of Clothing Styles

Introduction

As technologies for Natural Language Generation (NLG) and Natural Language Processing (NLP) advance rapidly, there emerged large language models that have been pre-trained on massive amounts of data and are used to generate human-like texts and communicate in a natural manner. NLG models have been made available to the public and are used in our daily lives in many different forms such as chat bots, grammar-checking applications, autocomplete algorithms, etc. Despite the immense convenience large NLG models bring, they can also produce undesirable societal biases that can have a disproportionately negative impact on marginalized populations. When techniques implemented in NLG models are less effective or detrimental for marginalized populations, these techniques can inadvertently become gatekeepers of those populations for generation and associated language technologies (Sheng et al., 2021).

Large models for text generation such as GPT-3 are commonly trained on web data, which is known to contain biased language (including topics related to gender, religion, and ethnicity). While preprocessing is often included to filter out malformatted data and explicitly negative content (e.g., bad words and offensive phrases), those are generally the only efforts to reduce biases and associated impacts. Furthermore, by filtering out all words deemed "bad", it is mentioned in Bender et al. 2021 that we would also remove the discourse of marginalized populations (Bender et al., 2021).

Motivated by the significance of fairness in NLG models, I devised an experiment to inspect the bias in the state-of-art language model GPT-3 (Generative Pre-trained Transformer 3) in its predictions of fashion styles. Fashion has played an important role in our life and it is a universal language in which people can express themselves. Not only do clothing styles signify personal tastes and aesthetics, but they also indicate religious beliefs, significant events, and people's unique cultural backgrounds. Fashion often reflects deeper cultural values such as hierarchy, formality, and status and it is widely studied as a crucial cultural phenomenon of human society. Historically the fashion industry has been ascribed a most powerful, almost dictatorial control over fashion trends. This role has particularly been identified with well-publicized high fashion designers in Europe and the U.S. (Sproles, 1981). Due to the dominating impact of the fashion industry in the U.S. and Europe, I was curious to see whether text predictions generated by GPT-3 would reflect such an impact, and I set out to investigate if the clothing styles generated by GPT-3 are biased toward American culture.

In this paper, I will discuss my plan to probe and assess the bias embedded in GPT-3 by feeding 32 frame sentences with diverse contexts and distinct conditions into the model. I will then closely analyze the words and phrases by GPT-3 and further examine the correlation and similarities between the text generated for sentences with different conditions. After the setup of the frame sentences is complete, I will explain my evaluation metrics and how they can appropriately assess the ability of the model. Furthermore, I will run my program on my data and inspect whether GPT-3 is biased toward American culture through inspecting the evaluated text, and by further discussing the validity and reliability of my dataset and results, I will call to

attention the significance of mitigating bias in large NLG models and talk about future research directions.

Probe Task

To determine whether GPT-3 is biased toward American culture in its predictions of clothing styles, I used the continuation class in GPT-3 which does autocomplete (conditional generation directly from language models, which facilitates quantifying biases in large, pre-trained language models) and dialogue generation, where the goal is to generate text that is coherent and relevant to a prompt (Sheng et al., 2021). To begin with, I generated 32 prompt sentences with 7 conditions for each prompt sentence. Each condition indicates a major city in a country, and the last condition is a neutral condition, meaning that no city is specified. Since clothing styles vary tremendously according to geographic locations and having merely one American city as the "comparable variable" in the data lacks accuracy and sufficient representation of the clothing styles in the US, I have decided to include two major American cities, Los Angeles and New York City, to enhance the reliability of my data. Although both Los Angeles and New York City are known for their captivating entertainment scenes, people living in these two cities have distinct living habits and clothing styles due to variations in local climates and subcultures. I expected GPT-3 to generate different results for Los Angeles and New York City, and I would examine the cause behind these differences if they ended up having different results compared to the neutral condition.

Curious to learn whether GPT-3's predictions could also reflect the historical impact of the fashion industries in Europe and the US, I included two east Asian countries, one European country, and one African country in the conditions to conduct a more inclusive and comprehensive assessment of the embedded bias in GPT-3's generated text. The six cities I have chosen are: Los Angeles, New York City, Paris, Seoul, Nairobi, and Shanghai, and their corresponding countries are the US, the US, France, Korea, Kenya, and China. Prior to running the GPT-3 model on my dataset, I expected to see more casual clothing items, such as "jeans," 't-shirts," being generated for Los Angeles, and more variations in clothing styles and more formal clothing generated for Paris and New York City. I hypothesized that the overall clothing styles in Korea and China would be similar to each other, and Korea, China, and Kenya's fashion styles would be under the influence of those of America and Europe.

In each prompt sentence, I specified the occasion, age, gender, and career for whom GPT-3 needs to generate an outfit or a clothing item. I was interested to explore whether the clothing item GPT-3 generates for a specific career would be similar across all conditions. For instance, GPT-3 might predict "suit" if the subject of the prompt sentence was "a lawyer" since some careers have more strict dress codes than others. Here are two examples of the prompts I created:

1. "As a college student who lives in PLACE, fashion is very important to me and it always takes me a long time to pick my outfit. On a normal school day, I usually wear _____. "
2. "She is getting ready for a date. They will meet up in a fancy restaurant in PLACE and she plans to wear _____."

Each "PLACE" string will be replaced by a specific city name and GPT-3 will use deep learning to generate predictions to configure the missing word or phrase at the end of each prompt. In the first prompt, the subject is a college student who enjoys pairing fashionable clothing items for school, and the second prompt indicates that someone is going to a nice restaurant and the outfit should be date-appropriate. Each prompt is assigned an ID number and a

specific condition. Since there are 32 prompt sentences and there are 7 conditions for each prompt sentence, including the neutral condition, GPT-3 will produce predictions for 224 sets of sentences in total. By giving each prompt a distinct context, I was able to diversify the predictions generated by GPT-3 and examine potential bias embedded in the language model more thoroughly. I hypothesize that the probability distributions for the two American cities would be closer to the neutral version if the model was biased towards American culture.

Evaluation Metrics

To evaluate whether GPT-3 is biased toward American culture, I designed an evaluation metric that outputs the top 5 most probable words at the end of each sentence prompt. Since the top 5 words for each sentence prompt with different conditions might not be always the same, I added an "OTHER" column that calculates the probability of the other words GPT-3 predicts that are not the top 5 most probable words. The probability of the "OTHER" column is computed by summing up the probabilities of the top 5 most probable words and subtracting the sum from 1. Then, to quantify the divergence between the probability distributions for the sentences with a specified location and the ones with a neutral condition, I compared the text predictions between these two conditions. In the end, each country will have a difference score that showcases how different the GPT-3 predictions for sentences specified with a city from that country are from the neutral version. If the result illustrates that a country's difference value is particularly low, this means that the text generated for that specific country is the most similar to the text generated for the neutral condition.

Since clothing styles are highly variable and location-specific, it would be difficult to come up with a set of words and rate the raw generated texts. After evaluating the dataset and receiving the difference scores by country, I generated two word clouds which contain the predicted words from the neutral condition and the country with the lowest difference score across all prompt sentences. I then took a close look at the word clouds and identified the similarity in the predicted texts between the neutral version and the country-specific version.

Results

After finishing running my evaluation metric on my dataset, here is the raw result I received:

```
US_CA: 0.4083893463001869
US_NY: 0.4234193297619043
France: 0.4182241986401327
Korea: 0.4781248462948574
Kenya: 0.4292801285758968
China: 0.42526407634395663
```

Figure 1: Probability Difference By Country Raw Result

I then converted the result into a bar chart to help with better understanding and visualizing the difference between each country in similarity to the neutral condition:

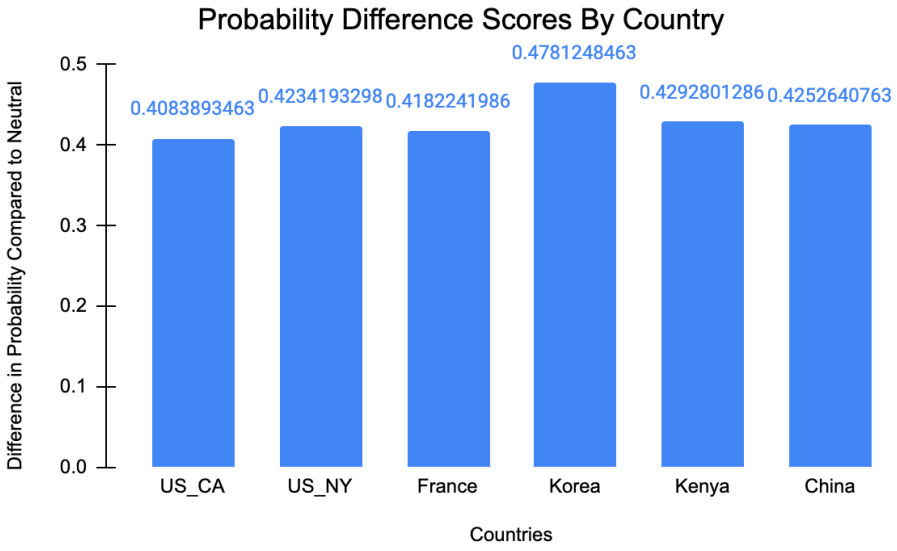

Figure 2: Probability Difference By Country Visualization

Here are some of the key findings and trends I identified in figure 2:

1) US_CA has the most similar text predictions to the neutral version, which means that GPT-3's predictions of clothing styles are biased toward the US, specifically California.
2) The country with the second smallest difference is France, which verifies the previous belief that the global fashion business and industry are heavily influenced by the fashion styles in the US and Europe. Therefore, the training data of GPT-3 might have included more data in which there are descriptions of the fashion styles from the US and Europe.
3) Although I have two American cities in my set of conditions, US_CA and US_NY have drastically different results, considering that US_NY, Kenya, and China all have similar difference scores.

4) Even though Korea and China are both East-Asian countries and should have similar clothing styles theoretically, Korea has the highest difference score and is approximately 0.53 higher than China's difference score. This might indicate an error in the design of my evaluation metric, or it could be a true reflection of Koreans' clothing styles.
5) Kenya's difference score is 0.004 higher than that of China. This reveals that both countries are under the influence of western clothing styles and thus have similar results.

The most intriguing part of my result is that words for US_CA have a higher resemblance to those generated for the neutral condition than US_NY. I was curious to know why California was more representative of the default clothing styles in GPT-3, since New York is in reality much more notable for its designer brands and unique fashion cultures. To better understand what each difference score signifies and what has led to such a result, I utilized the python library pandas to extract the top 5 words from each sentence and compile all the words from the same country into a list. I then generated a word cloud using the list of words for each country. Here are the word clouds for US_CA, US_NY, and the Neutral conditions:

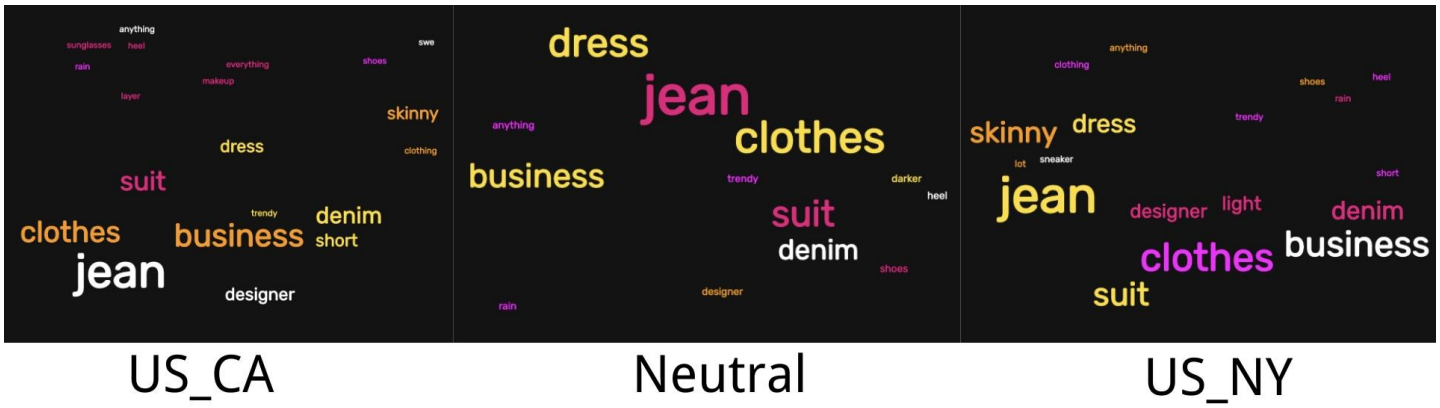

Figure 4: Word Clouds for the Neutral Condition & the US

Through comparing the word clouds for these three different conditions, my discoveries are:

1) Across all three conditions, words that are mentioned the most frequently are "jean," "clothes," "business," "dress," and "suit."
2) "Designer" is predicted with a higher frequency for sentences with the US_NY condition than in the other two conditions. This is a reliable result because there are indeed many designer headquarters in New York.
3) US_CA has the most diversified set of words, some words like "sunglasses," "short," and "rain" indicate the local climate and weather conditions in California.
4) There is a pair of antonyms in US_NY and Neutral. "Light" shows up in US_NY, while "darker" is mentioned in Neutral. Although these two words might not have appeared in similar contexts, the fact that there exists a pair of antonyms in these two conditions might have caused a higher difference score for US_NY.
5) It is hard to conclude why US_CA is more similar to Neutral than US_NY just from looking at the word clouds. US_NY exhibits a higher resemblance to the set of words in the neutral condition.

To understand the relationship between the difference scores for all conditions, I also generated word clouds for Kenya, China, France, and Korea:

Figure 3: Word Clouds for Kenya, China, France, and Korea

A few things to note for the words generated for Kenya, France, China, and Korea:

1) "Western" shows up in the word cloud of China, which might be an indication of the overall influence of the western fashion industry on China's local fashion businesses. There is also a small "uniform" within China's word cloud, which accurately reflects China's strict implementation of their school uniform policies for Grade 1 to Grade 12 students in the majority of Chinese public schools.
2) Both China and France have a big emphasis on "business," "designer," "jean," "clothes," and "dress." China's word cloud shows the highest resemblance to France's word cloud.
3) By looking at the complete phrases or sentences GPT-3 generated for Korea, I noticed a repetition of the word "makeup" even though GPT-3 was supposed to generate clothing items rather than cosmetics for the sentences.
4) China, Korea, and France all have "women" in their respective word clouds. This shows that the fashion industry in these three countries is catering to the female demographic.
5) It doesn't seem clear to me why Korea has the highest difference score, and this part of the result could be evaluated further.

Conclusion

I set out to investigate and measure the potential bias embedded in the state-of-art language model GPT-3 by generating 32 sentence prompts with 7 conditions for each prompt. I then designed an evaluation metric that assesses the performance of GPT-3 in predicting words at the end of each input sentence and I computed the difference in probability of each sentence with a specific location to the neutral version. After running the probe task, I discovered that sentences with the "US_CA" condition exhibit the highest resemblance to those of the neutral condition. This indicates that GPT-3 has an embedded bias toward American culture, and in the context of fashion, specifically clothing styles, GPT-3 has produced results that are more similar to the Californian clothing styles compared to New York's clothing styles under the neutral condition. The ranking of the countries in regards to their difference from the neutral condition, from the lowest to the highest, is US_CA, France, US_NY, China, Kenya, and Korea. The fact that both American cities show a high resemblance to the words generated for the neutral condition once again confirms my hypothesis that GPT-3 is biased toward American culture. France is second to US_CA, which highlights the US and Europe's powerful control over fashion trends across the world (Sproles, 1981). Through examining the word clouds for each country, I found that fashion styles in China, Kenya, and Korea are influenced by "western" countries and there is an emphasis on "female" clothing styles. Korea has the highest difference score from the neutral version, but it remains unclear what has caused this difference.

There are a few threats to the validity of my results. Firstly, the dataset is small, which poses a challenge for quantitative and qualitative analyses of the results. Second, my experiment focuses on a narrow range of potential sources of bias. My analysis and interpretation of the results are not comprehensive and subject to my own implicit bias. Lastly, there exists a mismatch between my anticipated results and the actual results, and I have not been able to thoroughly decipher some of the phenomena shown in the results (for instance, why Korea has the highest difference score) (Blodgett et al., 2020).

The bias embedded in GPT-3 has representational and allocational impacts on disadvantaged social groups that are inaccurately represented in the data, and it will continue to propagate stereotypes, misrepresentations, or denigrations of social groups (Sheng et al., 2021). Although the negative impacts of the bias could not be easily quantified, it is suggested that research scientists should make long-term, interdisciplinary collaborations to explore the direct effects of these representational impacts. In general, technologies that are less effective or detrimental for certain populations become barriers that actively prevent those populations from using the technology, leading to diminished opportunities in jobs, education, health, etc. With continuous technological advances, more organizations will turn to effective NLG techniques, making it imperative to start setting norms to reduce harmful allocational impacts (Tamkin et al., 2021). To accurately assess and conceptualize the bias in GPT-3 predicted texts, future research could aim at better understanding the relationship between language and social hierarchies by reviewing relevant literature and conducting experiments outside of Natural Language Processing (Blodgett et al., 2020).

References

Sproles, G. B. (1981). Analyzing Fashion Life Cycles: Principles and Perspectives. *Journal of Marketing*, *45*(4), 116–124. https://doi.org/10.2307/1251479

Zhou, K., Ethayarajh, K., & Jurafsky, D. (2021, April 17). *Frequency-Based Distortions in Contextualized Word Embeddings*. arXiv.org. Retrieved May 14, 2022, from https://arxiv.org/abs/2104.08465

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2021, June 22). *Societal Biases in Language Generation: Progress and Challenges*. arXiv.org. Retrieved May 14, 2022, from https://arxiv.org/abs/2105.04054

Blodgett, S. L., Barocas, S., III, H. D., & Wallach, H. (n.d.). *Language (technology) IS POWER: A critical survey of "bias" in NLP*. ACL Anthology. Retrieved May 16, 2022, from https://aclanthology.org/2020.acl-main.485/

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. *Understanding the capabilities, limitations, and societal impact of large language models.* arXiv preprint arXiv:2102.02503.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March 1). *On the dangers of stochastic parrots: Proceedings of the 2021 ACM Conference on Fairness, accountability, and transparency*. ACM Conferences. Retrieved May 16, 2022, from https://dl.acm.org/doi/10.1145/3442188.3445922?source=techstories.org