Abstract

Little attention is placed on analyzing nationality bias in language models, especially when nationality is highly used as a factor in increasing the performance of social NLP models. This paper examines how a text generation model, GPT-2, accentuates pre-existing societal biases about country-based demonyms. We generate stories using GPT-2 for various nationalities and use sensitivity analysis to explore how the number of internet users and the country’s economic status impacts the sentiment of the stories. To reduce the propagation of biases through large language models (LLMs), we explore the debiasing method of adversarial triggering. Our results show that GPT-2 demonstrates significant bias against countries with lower internet users, and adversarial triggering effectively reduces the same.

1 Introduction

Language models learn the context of a word based on other words present around it (Caliskan et al., 2017), and training an enormous dataset leads to the model learning powerful linguistic associations, allowing them to perform well without fine-tuning (Abid et al., 2021). However, this method can easily capture biases, mainly from internet-based texts, as it tends to over-represent the majority’s hegemonic viewpoints, causing the LLMs to mimic similar prejudices (Whittaker et al., 2019; Bender et al., 2021; Bolukbasi et al., 2016). Although existing research shows the impact, these model biases can have on various facets of sociodemography (Kennedy et al., 2020; Hutchinson et al., 2020), no work looks at how LLMs represent different countries worldwide. Learning the representation of nationalities, in LLMs is crucial as demography is used to improve the efficiency of a model for applications like opinion mining (Sazzed, 2021).

Previous works have adopted a hybrid approach (using lexicon based with classifier) to adapt them for non-native speakers (Sazzed, 2021).

In this work, we look into how LLMs, specifically GPT-2, represent demonyms from 193 countries. An example of potential bias in GPT-2 can be seen in Table 1. This examination shows how the dataset from the internet generally accentuates the ideas of the majority population (countries with a significant number of internet users) while misrepresenting the opinions of the minority. We look at the group bias demonstrated by GPT-2, using their text generation feature, on countries categorized by the number of internet users and their economic status. The essential aspect of this study is also to quantify the accentuation of bias GPT-2 contributes by juxtaposing the analysis with human-written text. Finally, we examine the potential solution of the group bias, in text generation models, by using the method of adversarial triggering where we positively trigger the prompts used by GPT-2 to provide better text.

2 Related Work

Research identifying bias in NLP models has shown that embedding models such as GloVe and Word2Vec, and context-aware dynamic embed-
dings, i.e., large language models (LLMs) such as BERT, automatically mimic biases related to gender (Kurita et al., 2019), race (Ousidhoum et al., 2021), disability (Venkit et al., 2022), and religion (Abid et al., 2021) from the language corpora used to train the model. The work done by Nadeem et al. (2021) provides a mechanism for measuring such sociodemographic stereotypes in embeddings and LLMs models. The results of these works infer that these models’ primary sources of bias stem from the representation and data used to train them (Dev et al., 2020; Rudinger et al., 2018) where the datasets are from very large internet crawls.

Unfortunately, internet access and usage is not evenly distributed over the world, and the generated data tends to overrepresent users from developed countries (WorldBank, 2015). Bender et al. (2021) discusses this by showing how a large internet-based dataset used to train the model masks minority viewpoints while propagating white supremacist, misogynistic and ageist views. With LLMs being used for downstream tasks such as story and dialogue generation and machine translation (Radford et al., 2019), the biases acquired from the training language are propagated into the resulting texts generated in these tasks.

Whittaker et al. (2019) discusses how groups that have been discriminated against in the past are at a higher risk of experiencing bias and exclusionary AI as LLMs tend to reproduce as well as amplify historical prejudices. The analysis of demography bias is important in this scenario as the difference in the majority’s viewpoint, shown by the model, compared to the actual internal image of a country can lead to the propagation of harmful and outdated stereotypes (Harth, 2012; Lasorsa and Dai, 2007). Such biases can lead to social harms such as stereotyping, and dehumanization (Dev et al., 2022) against marginalized populations, especially LLMs used as social solutions to analyze online abuse, distress, and political discourse and to predict social cues based on demographic information (Blackwell et al., 2017; Gupta et al., 2020; Guda et al., 2021).

3 Methodology

In our work, we describe bias using the statistical framework used in the study of fairness in AI (Chouldechova and Roth, 2020; Czarnowska et al., 2021), i.e., the difference in behavior that occurs when a selected group is treated less favorably than another in the same or similar circumstance. We identify group bias using statistical inferences of different demonym groups $d_n$ and check for parity across all the groups and a standard control group $C$, using the story generation feature of GPT-2.

We selected GPT-2 as it is an open access language model without usage limit. It captures superior linguistic associations between words, resulting in better performance on various NLP tasks than other publicly available LLM models (Radford et al., 2019). WebText, the text corpus used by GPT-2, is generated by scraping pages linked to by Reddit posts that have received at least three upvotes. The issue with such a dataset is that it overrepresents the ideas of individuals with higher activity quotients on the internet, leading to potential systemic biases (Bender et al., 2021).

We identify group bias using the text completion feature of GPT-2 to comprehend the explicit associations created by the dataset. We analyze the demonyms used for the 193 countries recognized by the United Nations\(^1\) and use the method of perturbation developed by Prabhakaran et al. (2019); Kurita et al. (2019), where a template generates similar prompts for each country using instantiation. We use the prompt $X$: \[ \text{[The <dem> people are]} \] and instantiate $<dem>$ with demonyms $d \in D$ (where $D$ is the set of 193 selected nationalities) to generate 100 unique\(^2\) stories per demonym, with a 500-word upper limit, using the GPT-2 API from Huggingface\(^3\). In order to generate the control $C$ and remove associations to any demonym, we generate 100 stories using the prompt \[ \text{[The people are]} \], resulting in a final corpus of 19,400 stories.

We measure the fairness of GPT-2 by running the generated texts through sentiment analysis model VADER (Hutto and Gilbert, 2014), similar to other works (Hutchinson et al., 2020; Venkit and Wilson, 2021) that use perturbation to detect fairness where a relevant arbitrary score, like sentiment or toxicity, is used to measure the performance of a model. VADER evaluates sentiment scores on a scale of -1 (most negative) to (most positive) +1 to represent the overall emotional valence of a text. Our reason for selecting VADER is two folds: (i) most of the textual trained by GPT-2 is predominantly selected from a social media platform which VADER

\(^1\)https://www.un.org/en/about-us/member-states
\(^2\)The authors of the paper manually examined 15 random stories generated for each prompt to make sure the texts generated were unique.
\(^3\)https://huggingface.co/GPT-2
Table 2: Analysis of most positive and negatively scored countries. \( f(\text{LLM}) \) denotes scores generated by GPT-2. \( f(\text{Hum}) \) denotes scores generated by non-AI text. \( f(\text{DeB}) \) denotes scores generated by post adversarial and \( \Delta f \) denotes bias accentuation.

<table>
<thead>
<tr>
<th>Demonym</th>
<th>Top Adjectives</th>
<th>( f(\text{LLM}) )</th>
<th>( f(\text{Hum}) )</th>
<th>( f(\text{DeB}) )</th>
<th>( \Delta f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>good, important, best, strong, true</td>
<td>0.375</td>
<td>0.301</td>
<td>0.672</td>
<td>0.126</td>
</tr>
<tr>
<td>Finland</td>
<td>good, important, better, free, happy</td>
<td>0.358</td>
<td>0.605</td>
<td>0.524</td>
<td>0.247</td>
</tr>
<tr>
<td>Ireland</td>
<td>important, good, better, difficult, proud</td>
<td>0.315</td>
<td>0.389</td>
<td>0.645</td>
<td>0.074</td>
</tr>
<tr>
<td>San Marino</td>
<td>good, important, strong, original, beautiful</td>
<td>0.314</td>
<td>0.577</td>
<td>0.649</td>
<td>0.263</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>good, important, legal, certain, better</td>
<td>0.287</td>
<td>0.102</td>
<td>0.372</td>
<td>-0.185</td>
</tr>
<tr>
<td>Libya</td>
<td>terrorist, clear, great, important, strong</td>
<td>-0.701</td>
<td>0.076</td>
<td>-0.055</td>
<td>0.777</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>important, affected, worst, difficult, dangerous</td>
<td>-0.702</td>
<td>0.232</td>
<td>0.079</td>
<td>0.934</td>
</tr>
<tr>
<td>Sudan</td>
<td>special, responsible, worst, poor, terrorist</td>
<td>-0.704</td>
<td>0.075</td>
<td>0.212</td>
<td>0.779</td>
</tr>
<tr>
<td>Tunisia</td>
<td>violent, terrorist, difficult, good, legal</td>
<td>-0.722</td>
<td>0.063</td>
<td>0.199</td>
<td>0.785</td>
</tr>
<tr>
<td>South Sudan</td>
<td>illegal, serious, dead, desperate, poor</td>
<td>-0.728</td>
<td>0.169</td>
<td>0.170</td>
<td>0.897</td>
</tr>
</tbody>
</table>

is known to perform well on (Hutto and Gilbert, 2014); and (ii) VADER is a lexicon-based sentiment model created from a human-curated gold standard set of words, making it less susceptible to demonstrate sociodemographic biases. We check this by running all 193 prompts \( f(DM) \) through VADER to identify explicit bias, but found none (as all scores were 0.00).

4 Results

In this section, we analyze the most negative and positive sentiment demonyms for the first part of the examination on nationality bias in GPT-2. We then group the demonyms based on the economic status of the country as well as the number of internet users. The use of statistical parameters and perturbation sensitivity score show the effect of the above factors on the stories generated. Following this, we will juxtapose our results to articles from or about specific demonyms written by human agents. Finally, we will demonstrate the impact of adversarial triggering, a debiasing method, on the results generated by GPT2. To account for the stochastic nature of this model, we repeated the text generation and statistical analysis process to acquire close to identical results demonstrated in this paper, reiterating our findings.

4.1 Analysis of Adjectives

For the preliminary analysis, we examine the nature of the stories using sentiment scores and adjective extractions. Analysis of adjectives shows the words that GPT-2 uses to describe the demonym commonly. Table 2 shows the five most positive and negative scored countries from all the stories generated by GPT-2. We use Textblob (Loria, 2018) to extract adjectives from the texts. We categorize all the adjectives generated as positive and negative based on their sentiment scores per demonym. Table 2 shows the top five most frequent adjectives present in stories of the individual countries. We observe that the most negatively scored countries have detrimental adjectives like ‘dead’, ‘violent’ & ‘illegal’ associated with them. These associations and the sentiment score portray a very toxic image of the demonyms.

4.2 Analysis of Internet Usage and Economic Status

We group the countries based on two factors, i.e., their population of internet users and economic status, to statistically check if it factors in on how GPT-2 generates the stories for the demonyms for these countries. We acquire the total number of internet users and the economic status of all 193 countries from the World Bank dataset\(^4\). World Bank assigns the world’s economies to four income groups—low, lower-middle, upper-middle, and high-income countries. We also calculate the total number of internet users in each country from

\(^4\)https://data.worldbank.org/
data collected by the World Bank on the internet usage parameters for all countries. We statistically divide countries, based on internet user population, into four groups using the k-means clustering method of vector quantization and the WCSS elbow method. The categorization of each country is present in our project repository.

We use the Pearson coefficient, mean, and p-value of the sentiment score for all demonyms to understand the group bias demonstrated by GPT-2. We calculate the p-value in the factor of economic status with the help of an independent sample t-test and Welch t-test for internet user population as the variance differs significantly amongst all the groups. Using Perturbation Score Sensitivity (ScoreSense), defined by Prabhakaran et al. (2019), we measure the extent to which a model prediction is ‘sensitive’ to specific demonyms. ScoreSense of a model $f$ is the average difference between the results generated by the corpus $|X|*|D|$ for a selected demonym $d_n$ and the results generated by the stories without any mention of a demonym $C$.

$$\text{ScoreSense} = \sum_{d_n \in D} [f(|X| \ast |d_n|) - f(C)]$$

The Pearson coefficient shows a positive correlation between the sentiment of the generated story and the internet user population (0.818), and the country’s economic status (0.935). Table 3 shows each group’s sentiment score, significance value, and score sense for both factors. Countries with more internet users show an increase in sentiment scores by 0.191 from the control group. On the other hand, scores for countries with low internet users dip by 0.124. We see similar behavior concerning economic status as well. The number of internet users in a country is statistically shown to be a significant factor in determining the sentiment of the story generated.

### 4.3 Evaluation of Human Written Stories

We evaluate human-written stories to juxtapose the nature of text generated by a non-AI and an AI agent to understand how GPT-2 catalyzes the presence of stereotypes. We randomly select 50 articles for each demonym, written about or from the selected country, from the NOW corpus (Davies, 2017), which contains data from 26 million texts written in English from online magazines and newspapers from various nations worldwide. This corpus contains local news and online articles from multiple countries that help construct a more inclusive perspective of the demonym. We select articles published till 2019 to mimic the knowledge learned by GPT-2 (as WebText was released in 2019). We depict the sentiment analysis acquired for all the stories, for a selected list of countries, in Table 2 through $f(\text{Hum})$. Comparing $f(\text{LLM})$ (sentiment scores of the text generated by GPT-2) to $f(\text{Hum})$, we can see that the overall sentiment score of stories generated by GPT-2 is more negative than the human-written articles.

We also notice countries like South Sudan and Sierra Leone, with a lesser $f(\text{Hum})$ value, receive a significantly negative score compared to countries that received an overall positive sentiment score. To understand this gap better, we define $\Delta f$ to measure negative bias accentuation caused by GPT-2 by measuring the difference between texts generated by non-AI and AI agent ($f(\text{Hum}) - f(\text{LLM})$). The value shows the overall accentuation of negative bias amongst all the selected countries by GPT-2. The score shows that lower countries (negative sentiment scores) are penalized substantially more ($\sim 0.834$) than top countries ($\sim 0.105$) concerning sentiment score. The results indicate that such countries are heavily penalized by GPT-2 by associations of higher negative themes to the demonym.

### 4.4 Debiasing using Adversarial Triggers

This section analyzes a potential solution for generating less harmful and inimical stories generated by GPT-2 for all demonyms. From our experimental
results in Table 2, we see that certain demonym
contain an unfavorable presence of toxic words that
can bring out a skewed perception of the country.
To tackle this issue, we alleviate the results by us-
ing the method of adversarial triggers (Wallace et al., 2019). For example, the prompt ‘French people
are’ can be changed to <positive adjective> French people are’ where <positive adjective> is
an adjective that adds a favorable context to the
demonym (eg: excellent, brilliant).

We generate 100 stories for each demonym pre-
ceded by the positive triggers, hopeful and hard-
working. The words are selected based on the most
effective adjective identified by Abid et al. (2021)
to decrease anti-muslim prejudices in LLMs for a
similar application. Table 2 and 4 show the results
obtained from debiasing. Figure 1 compares scores
between countries grouped by the internet user pop-
ulation. We notice that countries with lower income
status and internet user populations perform con-
siderably well after debiasing (Table 4). We also
see countries grouped as ‘High’ score lesser after
debiasing. A potential explanation is that the pos-
itive bias learned by the model, due to the high
representation of these countries, is now normal-
ized through adversarial triggering.

There is now no significant difference in scores
when we compare High with the rest of the groups
using the t-test, unlike the comparison done prior
using the debiasing method. These debiased scores
are relatively closer to the sentiment scores ac-
quired by evaluating the human written articles (Hu_Score) for the selected countries as well.

5 Discussion and Conclusion

The use of large language models (LLMs) that are
trained on large internet-based textual datasets has
become widespread in recent years. These mod-
els aim for scalability and universal solutions, but
in the process, biases towards potentially sensitive
words such as demonym can emerge. Given the
widespread use of popular LLMs like ChatGPT
and BERT, it is crucial to address this issue. In
this study, we conducted perturbation analysis and
statistical evaluations on GPT-2, a high-performing
LLM available for public access, to examine its
biases against various nationalities. Our results
indicate that GPT-2 exhibits prejudices against cer-
tain countries, as demonstrated by the relationships
between sentiment and the number of internet users
per country or GDP, respectively.

One potential cause of these demonym-based bi-
as is the large internet-based textual datasets used
to train the LLM, which tends to over-represent a
majority viewpoint while under-representing other
perspectives. Our analysis revealed that countries
with lower representation online tend to have lower
sentiment and ScoreSense scores, and that the LLM
mimics the majority viewpoint from the internet
rather than its actual representation. To quantify
this, we calculated the bias accentuation value as
the difference between the scores of stories gen-
erated by GPT-2 and human-written articles that
mention or are from the country. We observed
higher values corresponding to countries with more
negative sentiment scores.

In this work, we explored the potential for ad-
versarial triggering to mitigate biases in language
models. Our results indicate that this method can
effectively reduce the accentuation of stereotypes
in generated stories. Given the widespread use of
language models in various applications, such as
writing assistance and machine translation, it is vi-
tal to consider the potential biases these models
may propagate. Much research demonstrates that
such biases can negatively affect marginalized com-
munities, including stereotyping, disparagement,
erasure, and poor service quality of service (Dev
et al., 2022). Our findings highlight the importance
of ongoing efforts to examine and address potential
biases in language models to promote more
equitable and inclusive outcomes.

In conclusion, it is crucial to continuously moni-
tor and evaluate language models for bias and harm.
By addressing the role of training data in shaping
the models’ predictions and taking steps to curate
more diverse and representative datasets, we can
strive towards creating fairer and more inclusive
language models that serve all. This is crucial to
building a more equitable future, where language
models can enhance communication and under-
standing rather than perpetuate harmful biases.
Limitations

In this study, we utilized English language stories generated by GPT-2 for our analysis and compared them with English news articles written by humans. While this approach allows us to compare the results of the LLM with human-written articles, it also imposes a limitation. Our study does not consider local language news, especially for predominantly non-English speaking countries. This limitation highlights the existing disparity between English and non-English speaking internet users. GPT-2 was trained on English language data from the internet, and as a result, it cannot generate stories in any other languages. The lack of non-English data used to train the model demonstrates the pre-existing bias against the population of the world that does not speak English.

Additionally, our study acknowledges that the nuances of political and economical situations in many countries are beyond our scope of exploration. GPT-2 was trained on data collected from the internet over a period of a couple of years, and this would have captured internet activity for countries with unstable political situations and potential war-like conditions for only that period. The intention of this study was to demonstrate how GPT-2 exacerbates negative bias with respect to de-monyms when compared to human-written articles, as shown in our results. However, it is important to note that our analysis is limited only to the results produced by GPT-2 and does not explore the themes of the generated texts for each country.

References


Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.


Mark Davies. 2017. The new 4.3 billion word now corpus, with 4–5 million words of data added every day. In The 9th International Corpus Linguistics Conference.


Steven Loria. 2018. textblob documentation. Release 0.15. 2.


WorldBank. 2015. Individuals using the internet (% of population) - united states.