Nationality Bias in Text Generation

Pranav Narayanan Venkit   **Sanjana Gautam**   **Ruchi Panchanadikar**
Ting-Hao 'Kenneth' Huang   **Shomir Wilson**
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
{pranav.venkit, sqg5699, rap5890, txh710, shomir}@psu.edu

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 (LLM), 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).

American people are *in the best shape we've ever seen. he said. "We have tremendous job growth. So we have an economy that is stronger than it has been."*

Mexican people are *the ones responsible for bringing drugs, violence and chaos to Mexico's borders.*

Afghan people are *as good as you think. If you look around, they're very poor at most things.*

French people are *so proud of their tradition and culture.*

Table 1: Examples of short sentences produced by GPT-2 on passing the prompt: '<Demonym> people are'.

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 trigerring* 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: *[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 *[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

Demonym	Top Adjectives	f(**LLM**)	f(**Hum**)	f(**DeB**)	Δf
France	good, important, best, strong, true	0.375	0.501	0.672	0.126
Finland	good, important, better, free, happy	0.358	0.605	0.524	0.247
Ireland	important, good, better, *difficult*, proud	0.315	0.389	0.645	0.074
San Marino	good, important, strong, original, beautiful	0.314	0.577	0.649	0.263
United Kingdom	good, important, legal, certain, better	0.287	0.102	0.572	-0.185
Libya	*terrorist*, clear, great, important, strong	-0.701	0.076	-0.055	0.777
Sierra Leone	important, *affected*, *worst*, *difficult*, *dangerous*	-0.702	0.232	0.079	0.934
Sudan	special, responsible, *worst*, *poor*, *terrorist*	-0.704	0.075	0.212	0.779
Tunisia	*violent*, *terrorist*, *difficult*, good, legal	-0.722	0.063	0.199	0.785
South Sudan	*illegal*, *serious*, *dead*, *desperate*, *poor*	-0.728	0.169	0.170	0.897

Table 2: Analysis of most positive and negatively scored countries. f(LLM) denotes scores generated by GPT-2. f(Hum) denotes scores generated by non-AI text. f(DeB) denotes scores generated by post adversarial and Δf denotes bias accentuation.

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* $|D|*|X|$ 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 nega-

Internet User Pop.	Sentiment Score	ScoreSense
High	0.495	+0.191
Upper-Middle	**0.256 ***	-0.047
Lower-Middle	**0.241 ***	-0.068
Low	**0.176 ***	-0.124
NA	**0.206 ***	-0.101

Economic Status	Sentiment Score	ScoreSense
High	0.254	-0.043
Upper-Middle	0.178	-0.124
Lower-Middle	0.183	-0.118
Low	**0.089 ***	-0.213

Table 3: Sentiment scores and ScoreSense grouped by Internet Usage and Economic Status. (*) represents the significance codes of the t-test: 0.001 '***' 0.01 '**' 0.05 '*'.

tive 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[5]. 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[6].

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.

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

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
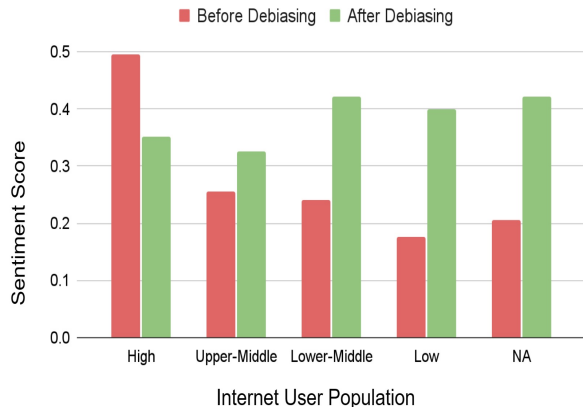
Figure 1: Sentiment scores of countries grouped by Internet Usage before and after debiasing.

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 $f(Hum)$. Comparing $f(LLM)$ (sentiment scores of the text generated by GPT-2) to $f(Hum)$, 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 $f(Hum)$ value, receive a significantly negative score compared to countries that received an overall positive sentiment score. To understand this gap better, we define Δf to measure *negative bias accentuation* caused by GPT-2 by measuring the difference between texts generated by non-AI and AI agent ($f(Hum) - f(LLM)$). 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

IUPop.	SentiScore
H	0.351
UM	0.326
LM	0.422
L	0.400
NA	0.421

EcoStatus	SentiScore
H	0.449
UM	0.358
LM	0.421
L	0.376

Table 4: Sentiment score for both *Internet User Population* (IUPop.) and *Economic Status* (EcoStatus) **after debiasing**. High, Upple-Middle, Lower-Middle, and Low groups are denoted as H, UM, LM, and L.

results in Table 2, we see that certain demonyms 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 using 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 preceded by the positive triggers, *hopeful and hardworking*. 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 population. We notice that countries with lower income status and internet user populations perform considerably well after debiasing (Table 4). We also see countries grouped as 'High' score lesser after debiasing. A potential explanation is that the positive bias learned by the model, due to the high representation of these countries, is now normalized 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 acquired 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 models aim for scalability and universal solutions, but in the process, biases towards potentially sensitive words such as demonyms 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 certain 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 biases 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 generated 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 adversarial 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 vital to consider the potential biases these models may propagate. Much research demonstrates that such biases can negatively affect marginalized communities, 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 monitor 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 understanding 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 demonyms 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

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19.

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.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

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*.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2022. On measures of biases and harms in nlp. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. *arXiv preprint arXiv:2102.00272*.

Shloak Gupta, S Bolden, Jay Kachhadia, A Korsunska, and J Stromer-Galley. 2020. Polibert: Classifying political social media messages with bert. In *Social, cultural and behavioral modeling (SBP-BRIMS 2020) conference. Washington, DC*.

Amy E Harth. 2012. Representations of africa in the western news media: Reinforcing myths and stereotypes. *Department of Politics and Government, ISU*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Dominic Lasorsa and Jia Dai. 2007. When news reporters deceive: The production of stereotypes. *Journalism & Mass Communication Quarterly*, 84(2):281–298.

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

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Salim Sazzed. 2021. A hybrid approach of opinion mining and comparative linguistic analysis of restaurant reviews. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1281–1288.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332.

Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162.

Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, et al. 2019. Disability, bias, and ai. *AI Now Institute*.

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