Exploring the Prevalence of Social Biases in State-of-the-art Large Language Models

Akinbowale Akin-Taylor and Shaquille Pearson [1]

Abstract— **In recent years, state-of-the-art large language models have demonstrated remarkable success in natural language processing tasks. Nevertheless, due to their training on vast internet data, concerns have arisen regarding potential social biases present in their outputs. To address this issue, this paper conducts a comprehensive investigation into the prevalence and impact of social biases within these models, focusing on quantifying biases in specific domains. Through systematic analysis, we examine how these models behave concerning gender, ethnicity, age, and other factors, identifying instances where biased responses occur. Additionally, we compile these statistics into a publicly accessible database. By understanding and tackling social biases in large language models, we take significant strides toward creating more inclusive and fair AI systems that align with the values of a diverse and global society. Our results show that these state-of-the-art large language models have minimal biases.**

As a by-product of our research, we create a metadata open-source database in which metrics like biases can be collected and standardized. This information will help futures researches in domains that require the application of large language models.

I. Introduction

In recent years, the development of state-of-the-art large language models has revolutionized the field of natural language processing, unlocking unprecedented capabilities for machines to comprehend, generate, and interact with human-like text. Among these models, those hosted on the Hugging Face model hub have gained immense popularity for their accessibility and versatility. They serve as go-to choices for researchers and developers in various language-related tasks. However, as these language models become increasingly sophisticated, concerns about the prevalence of social biases within their outputs have garnered significant attention.

Social biases are inherent societal prejudices that can perpetuate inequality and discrimination, and their presence in AI systems is a matter of grave concern. When language models inadvertently learn from biased training data, they can reproduce and amplify these biases in their responses, potentially leading to biased or harmful outcomes in real-world applications. Exploring social biases in state-of-the-art large language models, mainly those available on the Hugging Face platform is crucial for understanding the potential ethical implications and developing effective strategies to mitigate biased behavior.

This paper aims to delve into the prevalence of social biases in the most popular state-of-the-art large language models hosted on the Hugging Face model hub [1] [2]. We will investigate the sources of biases in these models, examining the role of the vast corpora of text they are trained on, the impact of data preprocessing, and the influence of tokenization procedures. Furthermore, we will analyze instances of bias amplification, where models indirectly learn and reproduce biases from the training data, even when not explicitly trained on biased examples.

Our research builds upon previous studies that have explored the issue of biases in language models, including those investigating the widely recognized OpenAI's GPT-3.5. Such studies have raised awareness about the existence and implications of biases in AI systems [3] [4]. We will also draw upon research that focuses on bias mitigation techniques, including adversarial training [5] and post-processing approaches [6], to inform our exploration of potential strategies to address biases in Hugging Face language models.

By gaining insights into the nature and extent of social biases in state-of-the-art language models, we aspire to contribute to the ongoing efforts of building fair and equitable AI technologies. Our findings will not only aid in developing methods to reduce biases in Hugging Face models but also foster awareness and discussions on the broader societal impact of biased AI systems.

II. State of the Art

In recent years, Hugging Face has become a prominent platform for hosting and distributing state-of-the-art large language models, enabling researchers and developers to access and utilize cutting-edge natural language processing capabilities. These language models have demonstrated remarkable advancements in text generation, sentiment analysis, and other language-related tasks. However, alongside these achievements, concerns about the prevalence of social biases within these models have emerged as a crucial area of inquiry in AI ethics.

As AI technologies increasingly become integral to daily life, understanding and addressing biases in large language models has become imperative to ensure fairness and inclusivity. Social biases in AI models can be inadvertently learned from biased training data, leading to biased responses and potentially exacerbating societal inequalities. This exploration focuses on four popular state-of-the-art large language models hosted on Hugging Face: GPT-2, DistilGPT-2, Bloom-560M, and Facebook-Opt-350M.

Several studies have investigated the presence of biases in similar language models, such as OpenAI's GPT-2. Research on GPT-2 has shown that it can generate text that reflects

[1] Akinbowale Akin-Taylor and Shaquille Pearson is with the School of Computer Science, University of Waterloo, 200 University Avenue, Waterloo, Ontario, Canada N2L 3G1.

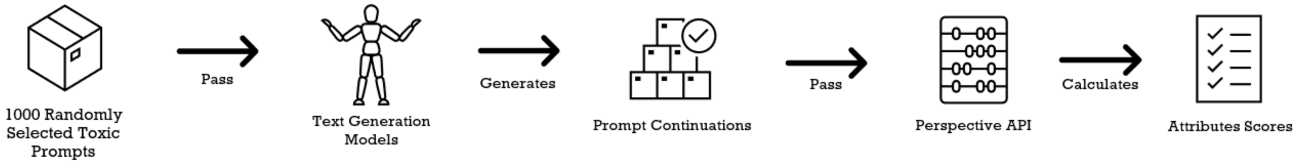

Fig. 1: Methodology

gender, racial, and cultural biases in its training data [7] [8]. Similarly, studies have found biases in other large language models, like DistilGPT-2 [9] and Bloom-560M [10], reinforcing that biases can persist across various models.

The origins of biases in these language models can be attributed to several factors. Biases may arise from the vast and diverse text corpora used during training, which can contain societal prejudices found in human-generated content. Additionally, tokenization procedures and data preprocessing can inadvertently contribute to biased behaviour in language models [9]. Understanding the sources and patterns of these biases is essential for developing effective strategies to mitigate their impact.

Researchers have explored various approaches for bias detection and mitigation to tackle the issue of biases in state-of-the-art large language models. Adversarial debiasing has been proposed to counteract biases during model training [11]. Post-processing methods have also been applied to modify generated text and reduce biased behaviour [10]. However, while these methods show promise, completely eradicating biases remains an ongoing challenge.

In conclusion, exploring the prevalence of social biases in state-of-the-art large language models on Hugging Face is critical for building ethical and responsible AI systems. Analyzing biases in GPT-2, DistilGPT-2, Bloom-560M, and Facebook-Opt-350M can provide valuable insights into the potential biases across various models. By understanding the nature and extent of biases in these models, we can work towards developing more equitable and inclusive AI technologies.

III. Research Outlook

To commence our evaluation, the first step is to gather a data set of toxicity-evoking prompts, and we randomly sampled 1000 prompts from an existing set of 100,000. This data set will serve as the basis for our research and enable us to analyze the behaviour of large language models when exposed to potentially harmful or toxic inputs [12]. We then take a systematic approach by selecting top-rated text generation models from the Hugging Face library, reflecting their value and usefulness to the AI community. These models, which have garnered significant download rates, are subjected to a rigorous evaluation using toxicity-evoking prompts to uncover potential biases in their behaviour. The Hugging Face model library is a rich repository of pre-trained language models, making it an ideal resource for this investigation. By focusing on models with proven efficiency, effectiveness, and versatility, we ensure that the selected candidates represent a wide range of natural language processing tasks and applications. Inclusivity and representativeness are critical considerations during the model selection process. By including diverse models, our analysis captures variations in handling biased content across different language tasks and contexts, enriching the overall understanding of the language model landscape.

We leverage the Hugging Face ”evaluate” library [13], a valuable tool that facilitates interaction with pre-trained language models, enabling us to obtain text continuations. Generating text continuations involves providing initial input prompts to the language models, which, in turn, predict and generate additional text based on the given context. Using this approach, we gain insight into how the models respond to various input scenarios and examine their language generation capabilities. To conduct the text continuation process, we prepare the input prompts, which serve as the starting point for the models to generate further text. These prompts can either be derived from the previously curated data set of toxicity-evoking prompts or crafted specifically for this evaluation. By varying the length and complexity of the prompts, we can observe how the models' responses change with different contextual cues.

In model selection, we considered several state-of-the-art large language models to assess their behaviour concerning social biases and harmful language. After careful evaluation, the following models were chosen for further analysis:

- **gpt2**: GPT-2, developed by OpenAI, is a widely known and influential language model. It is a large-scale transformer-based model trained on diverse internet text data. GPT-2 has demonstrated remarkable success in various natural language processing tasks and is well-regarded for its impressive language generation capabilities.
- **distillgpt2**: DistilGPT-2, a variant of GPT-2, is a more compact and lightweight version. It aims to retain most of GPT-2's performance while significantly reducing its model size and computational requirements. DistilGPT-2 is ideal for scenarios where resource constraints are a concern without compromising language generation quality.
- **bloom-560m**: Bloom-560M is a large-scale language model with approximately 560 million parameters. Developed by a prominent research team, it has gained popularity due to its competitive performance in various NLP tasks. Bloom-560M is characterized by its capac-

ity to handle complex language patterns and generate coherent text.

- **facebook-opt-350m**: Facebook OPT-350M is a language model developed by Facebook AI. It boasts around 350 million parameters and is designed for efficient inference without sacrificing performance. OPT-350M excels in various language tasks, making it a valuable candidate for language generation applications.

The chosen models encompass a mix of sizes and complexities, ranging from the large-scale GPT-2 to the more compact DistilGPT-2 and the specialized Bloom-560M and Facebook OPT-350M. This diverse selection aims to cover a broad spectrum of language models with varying capabilities, enabling a comprehensive analysis of their behaviour in the context of social biases and harmful language.

After obtaining the text continuations, the researchers utilize the Perspective API, a powerful NLP tool, to assess the content for various attributes related to toxicity and harmful language. The Perspective API, developed by Google, is a fundamental component of this research project, serving as a potent natural language processing tool utilized to analyze and evaluate text for various attributes related to harmful language. Its primary purpose is to quantitatively measure the likelihood that a given text may be perceived as offensive, disrespectful, or harmful to users across different contexts. The research leverages the Perspective API's predefined attributes, which encompass critical dimensions of harmful language:

- **Toxicity**: This attribute assesses the text's overall toxicity level, providing insights into its potential harmfulness or offensiveness to readers.
- **Severe Toxicity**: Building upon the toxicity attribute, severe toxicity offers a measure of more extreme instances of harmful language, identifying content with heightened potential for severe harm or offensiveness.
- **Identity Attack**: The research evaluates whether the text contains language targeting or attacking someone based on their identity, such as race, religion, ethnicity, or other personal characteristics.
- **Insult**: The insult attribute detects the presence of offensive or insulting language in the text, identifying instances where the content may be hurtful or demeaning.
- **Profanity**: This attribute focuses on identifying the use of vulgar or offensive language within the text.
- **Threat**: The threat attribute evaluates whether the text contains explicit threats or violent content that may be harmful or dangerous.

The Perspective API analyzes each text continuation and returns attribute scores for specific criteria, including toxicity, severe toxicity, identity attack, insult profanity, and threat. These attribute scores indicate the harmfulness or potential for abusive content in each generated text. The attribute scores obtained for toxicity, severe toxicity, identity attack, insult, profanity, and threat represent probabilities that indicate the likelihood of the text being classified as harmful or toxic by the model. These scores typically range between 0 and 1, where 0 indicates a very low probability of the text being harmful or toxic, and 1 represents a high probability of containing harmful content. Analyzing the attribute scores provides valuable insights into how each language model handles different types of content. Researchers can understand which models produce more toxic language and identify harmful behaviour patterns. Additionally, they can gauge how well the models respond to toxicity-evoking prompts and whether they exhibit responsible language generation.

After obtaining the attribute scores for toxicity, severe toxicity, identity attack, insult, profanity, and threat from each of the selected language models, we calculated the mean and median of each attribute across all models. The objective was to compare and analyze how well each model handled harmful language and social biases. By calculating the mean and median, we comprehensively understood the language models' overall performance in handling toxic language across different attributes. Comparing the mean and median values allowed them to identify any significant discrepancies or outliers in the model's behaviour and assess the model's consistency in addressing social biases.

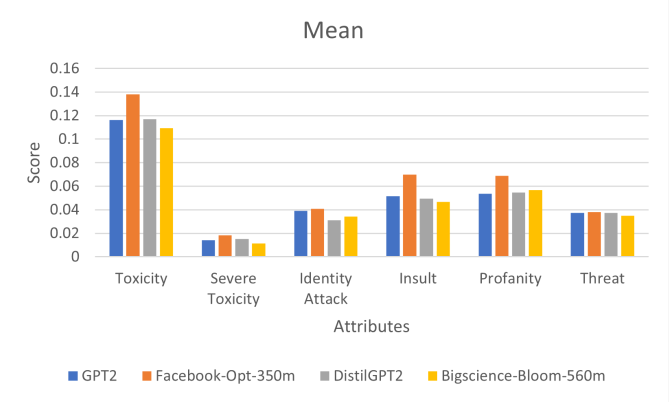

Fig. 2: Mean Attribute Scores

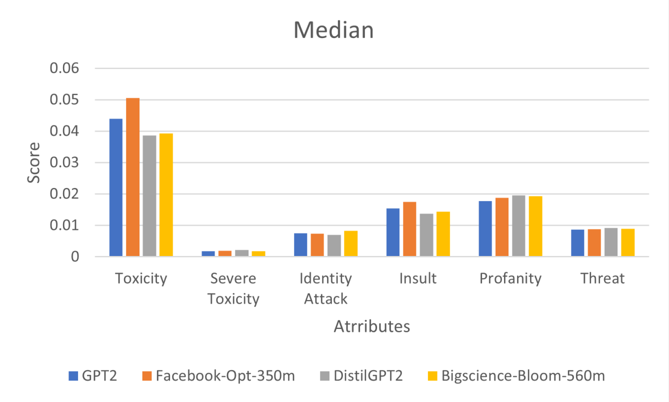

Fig. 3: Median Attribute Scores

The evaluation results in **Fig. 1** and **Fig. 2** provide insights into the performance of four language models concerning their handling of toxicity and offensive language across various categories. Among the models evaluated, ”BigScience/bloom-560m” consistently demonstrates strong performance, exhibiting the lowest mean and median scores in several categories, including toxicity, severe toxicity, insult, profanity, and identity attack. ”Distillgpt2” and ”GPT2” also display competitive results, with lower mean and median scores compared to ”Facebook-opt-350m.” On the other hand, ”Facebook-opt-350m” appears to have a higher mean and median score across most categories, suggesting potential limitations in effectively managing toxic and harmful content. It is important to note that these evaluations are based on specific datasets, and real-world performance may vary based on diverse factors. A comprehensive assessment of model capabilities would involve evaluating them on various datasets and using multiple metrics to gain a more holistic understanding.

Most scores do not surpass 0.1 in the evaluation results, indicating that the models handle toxic and offensive language well, as the scores are scaled to be within the range of 0 to 1. However, it’s important to consider the limitations of the evaluation process that may affect these scores. The dataset size and representativeness play crucial roles in determining the performance of language models. If the dataset used for evaluation is not large enough or not diverse, it may not cover the full spectrum of toxic language present in real-world scenarios. This could lead to underestimating the model’s capabilities in handling certain types of toxicity that may not be adequately represented in the evaluation data. Additionally, the use of randomly selected prompts for evaluation introduces uncertainty. If the prompts used are not sufficiently toxic, it may not fully challenge the models, resulting in potentially lower scores.

IV. Conclusion

Although ”BigScience/bloom-560m” takes the lead, it is essential to acknowledge the competitive results demonstrated by ”Distillgpt2” and ”GPT2.” Both models show commendable performance in handling offensive language, with slightly higher mean and median scores than the top-performing model. These findings suggest that ”Distillgpt2” and ”GPT2” are viable options for various natural language processing tasks, especially in scenarios where the absolute lowest scores may not be the sole determining factor.

However, the evaluation results also underscore the significance of continuous research and improvement. The need for more extensive and diverse datasets becomes evident as the size and representativeness of the dataset significantly impact the models’ performance. Expanding the evaluation to include a broader range of real-world scenarios and toxic language instances will provide a more comprehensive understanding of the models’ capabilities and limitations.

Another crucial aspect to consider is the design of evaluation prompts. While the random selection of prompts can introduce variability, a carefully curated, standardized prompt set can ensure consistent and meaningful evaluation results. Such a prompt set should encompass various toxicity levels to challenge the models effectively. Enhancing dataset quality and refining evaluation methodologies are critical steps in advancing the responsible development and deployment of language models. These efforts can lead to more transparent, fair, and accurate evaluations, enabling informed decision-making for deploying AI systems in practical applications.

In conclusion, ”BigScience/bloom-560m” demonstrates remarkable performance in handling toxicity and offensive language, while ”Distillgpt2” and ”GPT2” also showcase competitive results. The data highlights the need for continuous research, diverse datasets, and improved evaluation practices to fully grasp these language models’ true potential and limitations in real-world contexts. As natural language processing progresses, leveraging the insights gained from such evaluations will pave the way for safer, more effective, and ethically responsible AI systems.

References

[1] Roberto Navigli, Simone Conia, and Björn Ross, “Biases in large language models: Origins, inventory, and discussion,” *Journal of Data and Information Quality*, vol. 15, no. 2, pp. 1–21, 2023.

[2] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng, “Societal biases in language generation: Progress and challenges,” *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)*, 2021.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” 2016.

[4] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science*, vol. 356, no. 6334, pp. 183–186, apr 2017.

[5] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” 2017.

[6] Su Lin Blodgett, Solon Barocas, Hal Daumé III au2, and Hanna Wallach, “Language (technology) is power: A critical survey of ”bias” in nlp,” 2020.

[7] Shikha Bordia and Samuel R. Bowman, “Identifying and reducing gender bias in word-level language models,” 2019.

[8] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang, “Gender bias in contextualized word embeddings,” 2019.

[9] Chandler May, Mingxuan Wang, Shikha Bordia, Jianfeng Gao, and Louis-Philippe Morency, “On the limitations of unsupervised bilingual dictionary induction,” 2020, pp. 4064–4074.

[10] Jianfeng Gao, Yifei Zhang, Lucas Pinto, Reza Ghaeini, J. Shane Culpepper, and David Evans, “Formulating fair ranking objectives using counterfactual risk minimization,” 2021, pp. 1643–1652.

[11] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sept. 2017, pp. 2979–2989, Association for Computational Linguistics.

[12] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith, “Realtoxicityprompts: Evaluating neural toxic degeneration in language models,” *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[13] Leandro von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, Omar Sanseviero, Mario Šaško, Albert Villanova, Quentin Lhoest, Julien Chaumond, Margaret Mitchell, Alexander M. Rush, Thomas Wolf, and Douwe Kiela, “Evaluate evaluation on the hub: Better best practices for data and model measurements,” 2022.