Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-4 Issue-2, February 2024

Bias in Text Generative Open AI

Sai Asrith Devisetti

Abstract: ***The rise of text generation models, especially those powered by advanced deep learning architectures like Open AI's GPT-3, has unquestionably transformed various natural language processing applications. However, these models have recently faced examination due to their inherent biases, often evident in the generated text. This paper critically examines the issue of bias in text generation models, exploring the challenges posed, the ethical implications it entails, and the potential strategies to mitigate bias. Firstly, we go through the causes of the origin of the bias, ways to minimize it, and mathematical representation of Bias.***

Keyword: Deep Learning, Generated Text, Ai

I. INTRODUCTION

Text-generative models have become most prominent in our lives, and we sometimes have too much faith in the outcomes. However, the responses that a text-generative AI model gives aren't neutral.

A. What is Bias?

[2] Bias here refers to the Statistical Bias. Statistical Bias describes the tendency of the generated output to be inappropriate or skewed from reality. The advanced use of artificial intelligent systems brings numerous problems and challenges for users. To ensure fair and reliable data, detecting and addressing bias is crucial, particularly in decision-making or machine learning.

- **Bias in Data:** Bias in data arises from how populations are sampled and defined, as well as how features or labels are chosen or collected.
- **Bias in Modeling:** Bias might arise during model iteration and evaluation. Also, there might be bias during model construction where distinct populations are inappropriately combined.

II. BACKGROUND & RELATED WORK

In recent years, the bias in Search Engines has come to light. Now, the text generative models are the most used and trusted platforms, so evaluating bias in them is essential. Bias is important because the returns must be balanced representation of all possible outcomes [9][10][11].

Below are some research papers that I have been through to understand the bias in Search Engines:

1. [5] Search Engine Bias and the Demise of Search Engine Utopianism: This paper has the algorithms that search engines use to make editorial choices. Also, how search engine editorial choices create Biases. Along with the ways to mitigate bias.
2. [7] Evaluation Metrics for Measuring Bias in Search Engine Results: The Search engine decides what we see for a given query. The keywords for the publication were "Bias evaluation", "Fair ranking", "Search Bias", and "Web Bias". This paper explains how to estimate bias from unknown algorithms and deals with some approximate algorithms used by search engines for providing the output [8].

A. Bias in Data

[3] In continuation of the Bias in Data, bias can show up in any form:

- **Historical Bias:** Historical bias is already an existing bias in the world that stepped into the data.
- **Representation Bias:** Representation bias arises from the way data is collected and defined.
- **Measurement Bias:** Measurement bias occurs when the features or labels used for training AI models are chosen in a way that introduces bias [12].

B. Bias in Modelling

[3] In continuation of the Bias in Modelling, how perfect the data might be, there are high chances for bias to show up while modelling:

- **Evaluation Bias:** Evaluation bias occurs when the evaluation metrics used to assess the performance of AI models are biased or do not consider all relevant factors.
- **Aggregation Bias:** Aggregation bias refers to the bias that can emerge when data is aggregated or summarized at different levels.

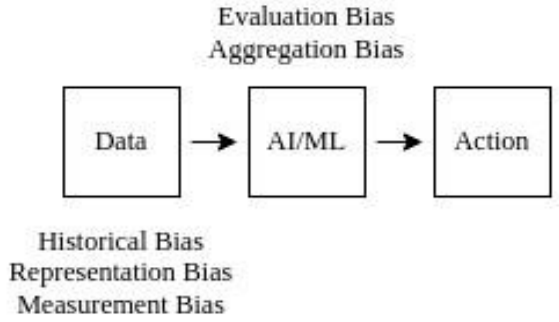

Figure 1: Origin of Bias

III. MECHANISMS FOR TEXT GENERATION IN AI MODELS

The data is stored in the main server and retrieved based on the user query. The following hierarchy is followed for generating the output:

Manuscript received on 08 January 2024 | Revised Manuscript received on 17 January 2024 | Manuscript Accepted on 15 February 2024 | Manuscript published on 30 May 2024.
**Correspondence Author(s)*
Sai Asrith Devisetti*, Department of Computer Science and Engineering International Institute of Information Technology Hyderabad (Telangana), India. E-mail: sai.asrith@research.iiit.ac.in, ORCID ID: 0009-0008-9691-8700

- **Crawling:** Crawling is considered to be the first step in data exploration, the AI models traverse through the available data and store the required data in the local database, as per the user query. The model in this stage is trained to comprehend the diverse linguistic landscape
- **Indexing:** The data stored in the local database, is structured and organized in the internal memory. In this stage, the model is trained to recognize patterns.
- **Ranking:** The structured and organized data will be ranked based on the user requirement and the highest ranked data will be outputted.

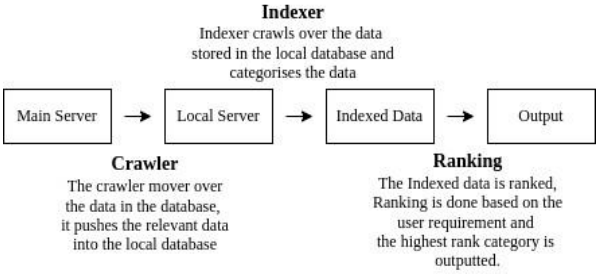

Figure 2: Hierarchy of Output Generation

IV. STRATEGIES FOR MITIGATING BIAS

Bias could be added from different stages in the mechanism used for text-generative AI models, we are not considering the Historical Bias as it is most common and could be prevented by some strategies, Representation bias arises in the crawling, Measurement bias arises in the Indexing, Evaluation bias in the Ranking, and Aggregation bias in the output. The total bias is the combination of the above deviations.

Although Bias can be mitigated following some preventing measures:

- **Diverse Training Data:** Incorporate diverse and representative training data that covers various demographic groups, cultures, and perspectives. This helps reduce bias originating from historical underrepresentation.
- **Data Preprocessing and Cleansing:** Implement rigorous data preprocessing and cleansing techniques to identify and mitigate biased data within the training data. This can involve removing or neutralizing biased examples and entities.

V. QUANTIFYING BIAS

Statistical Bias in Estimation

Statistical bias is a characteristic of a statistical technique or its outcomes where the expected value of the results differs from the true underlying quantitative parameter being estimated. Bias is defined as follows: Let T be a statistic used to estimate a parameter θ, and let E(T) denote the expected value of T. Then,

$$\text{bias}(T,\theta) = \text{bias}(T) = \text{E}(T) - \theta$$

is termed the bias of the statistic T concerning θ. If bias(T,θ) = 0, then T is considered an unbiased estimator of θ; otherwise, it is considered a biased estimator of θ.

For Example, while calculating the system's bias, T can be specific text generated by the AI model, and θ can be ideal or unbiased text. Thus, calculated bias represents how much the generated text differs on average, from the ideal, unbiased text.

Through the process, we work on minimizing the bias by exploring various causes for bias.

Although the algorithms for the output generation and feature selection are not transparent, there are approximate algorithm auditing technologies that provide an effective means for evaluating bias. For example, the bias in the content can be estimated using the Crowd Sourcing technique [6].

Let us consider S to be the set of crawling bots and Q be the set of queries. Let s ϵ S, and q ϵ Q, then bias at each level can be written as:

$$\beta_f(r) = f_o(r) - f_e(r)$$

where f is a function that measures the likelihood of r in satisfying the information need of the user about the view Output(o) and the view Estimate (e),. We note that ideological bias is measured in the same way by transforming the stances of the documents into ideological leanings. Before defining f, from Eq. (1), we define the mean bias (MB) of a search engine s as:

$$\text{MB}_f(s,Q) = \frac{1}{|Q|}\sum_{q\in Q}\beta_f(s(q))$$

An unbiased model would produce a mean bias of 0. A limitation of MB is that if a generative model is biased towards the Output(o) view on one topic and bias towards the Estimate(e) view on another topic, these two contributions will cancel each other out. In order to avoid this limitation, we also define the mean absolute bias (MAB), which consists in taking the absolute value of the bias for each r. Formally, this is defined as follows:

$$\text{MAB}_f(s,Q) = \frac{1}{|Q|}\sum_{q\in Q}|\beta_f(s(q))|$$

VI. CONCLUSION AND FURTHER WORKS

In this work we proposed a new representation of the total bias for the Generative AI model also we examined the origin of bias and precautions to prevent the Bias

We did not go through the Bias Evaluation and Statistical analysis, i.e, what factors influence the bias most. This could be left for future progress [1][4].

DECLARATION STATEMENT

Funding	No, I did not receive.
Conflicts of Interest	No conflicts of interest to the best of my knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Not required.
Authors Contributions	I am only the sole author of the article.

REFERENCES

1. On the Apparent Conflict Between Individual and Group Fairness.
2. Understanding Bias and Fairness in AI Systems.
3. Managing Bias in Machine Learning.
4. Confirmation Bias: Roles of Search Engines and Search Contexts.

5. Search Engine Bias and the Demise of Search Engine Utopianism.
6. Bias in Search Engines and Algorithms.
7. Evaluation Metrics for Measuring Bias in Search Engine Results.
8. Kanani, P., & Padole, Dr. M. (2019). Deep Learning to Detect Skin Cancer using Google Colab. In International Journal of Engineering and Advanced Technology (Vol. 8, Issue 6, pp. 2176–2183). https://doi.org/10.35940/ijeat.f8587.088619
9. Text Generation using Neural Models. (2019). In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 2S, pp. 19–23). https://doi.org/10.35940/ijitee.b1006.1292s19
10. Chellatamilan, T., Valarmathi, B., & Santhi, K. (2020). Research Trends on Deep Transformation Neural Models for Text Analysis in NLP Applications. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 9, Issue 2, pp. 750–758). https://doi.org/10.35940/ijrte.b3838.079220
11. Mathew, S. (2024). An Overview of Text to Visual Generation Using GAN. In Indian Journal of Image Processing and Recognition (Vol. 4, Issue 3, pp. 1–9). https://doi.org/10.54105/ijipr.a8041.04030424
12. Radhamani, V., & Dalin, G. (2019). Significance of Artificial Intelligence and Machine Learning Techniques in Smart Cloud Computing: A Review. In International Journal of Soft Computing and Engineering (Vol. 9, Issue 3, pp. 1–7). https://doi.org/10.35940/ijsce.c3265.099319

AUTHOR PROFILE

Sai Asrith Devisetti is a burgeoning researcher and technologist whose academic and professional journey epitomizes dedication, innovation, and a commitment to excellence. Sai Asrith's academic excellence is highlighted by his achievement in the Indian National English Olympiad, where he secured an impressive All-India Rank of 5. Currently pursuing a Bachelor of Technology in Computer Science and Engineering at the prestigious International Institute of Information Technology, Hyderabad (IIIT Hyderabad), Sai Asrith is poised to make significant contributions to the field of artificial intelligence and machine learning. IIIT H's rigorous academic environment has equipped him with a robust understanding of core computer science principles, encompassing areas such as data structures, algorithms, operating systems, computer networks, and database management systems. His coursework in machine learning, discrete mathematics, probability and statistics, quantum computation, and linear algebra further underscores his comprehensive and multidisciplinary approach to computer science. Sai Asrith's professional experience is characterized by a series of impactful internships and projects that demonstrate his ability to translate theoretical knowledge into practical applications. His tenure at the Raj Reddy Centre, working on the GradeMate project, involved the development of an application that assesses the correctness of spoken answers using automatic speech recognition (ASR) models and word comparators. In his recent paper on" Bias in Text Generative AI," Sai Asrith delves into the critical issue of inherent biases in AI models developed by leading organizations like OpenAI. His research explores how these biases manifest in generative text outputs and the implications they have on fairness and ethics in AI applications. By examining various case studies and employing rigorous analytical methods, he aims to identify root causes and propose strategies to mitigate these biases. His work not only contributes to the academic discourse on AI ethics but also aims to influence the development of more equitable AI systems.