Detecting unknown Biases in Bios

Théodore Fougereux - Tom Fizycki

13 novembre 2024

Table des matières

1 Unknown bias and bios dataset

a. The Bias in bios dataset

The *Bias in Bios* dataset was methodically curated to investigate gender bias within a myriad of professional contexts, with the aim of unraveling nuanced patterns in the portrayal of individuals. This comprehensive dataset, sourced from online biographies written in English, provides valuable information such as names, pronouns, and occupations, enabling researchers to delve into the intricacies of gender bias across a diverse spectrum of professions.

In constructing the dataset, the researchers identified and categorized the twenty-eight most frequent occupations, offering a granular view of bias prevalence within each. Occupations like professors emerged as highly represented, boasting 118,400 biographies out of 397,340, while others, like rappers, exhibited lower frequencies with only 1,406 biographies. This occupation-centric approach enhances the dataset's utility, allowing researchers to explore and compare bias trends across different professional landscapes.

This dataset has been previously explored in [6]. The focus of the prior study was on biases encountered in predicting job roles mentioned in the biographies. However, the biases investigated were predetermined (e.g., gender, sexual orientation, origins, etc.) and correctable using methods outlined in [18], for instance. Consequently, the exploration of methods to identify existing biases a priori remains an open question.

b. Main characteristics

Key characteristics of the *Bias in Bios* dataset contribute to its richness and applicability :

1. **Occupational Diversity :** The dataset's inclusion of a broad array of occupations fosters a holistic examination of gender bias. This diversity empowers researchers to discern how bias manifests uniquely in various professional domains, offering insights into societal perceptions and stereotypes.
2. **Biography Length Variation :** Variability in biography length, ranging from concise eighteen-token profiles to detailed 194-token narratives, adds complexity to the dataset. Researchers can explore how the length of biographies may influence the prevalence and perception of bias, contributing to a more nuanced understanding of biased representations.
3. **Demographic Considerations :** Acknowledging potential discrepancies between the demographics of individuals in online biographies and the overall workforce is crucial. Researchers must interpret findings with caution, recognizing the dataset's limitations in representing the entire spectrum of workforce demographics.
4. **Dataset Limitations :** The creators transparently communicate that *Bias in Bios* does not capture all available online biographies. This transparency prompts researchers to consider the dataset's representational scope, urging caution in generalizing findings beyond its confines. Moreover, some of the scraped biographies seem professionally irrelevant or provocateur, hence raising some concerns on the reliability of the dataset.

c. Main usecases of the Bias in Bios dataset

Potential uses and applications of the *Bias in Bios* dataset extend across various domains :

1. **Algorithmic Fairness Research :** The dataset serves as a benchmark for developing and evaluating algorithms aimed at detecting and mitigating gender bias in biographical data, contributing to advancements in algorithmic fairness.
2. **Societal Studies and Policy Formulation :** Scholars in social sciences can leverage the dataset to conduct studies on societal attitudes and stereotypes, informing policy initiatives that aim to mitigate biases in professional representation.
3. **Educational Resources :** The dataset can be employed in educational settings to raise awareness about bias, allowing educators to incorporate real-world examples into discussions on diversity and inclusion.
4. **Corporate Diversity Initiatives :** Organizations can use the dataset to assess biases in their communication materials and automatic recruiting processes, aiding in the development of strategies to promote diversity, equity, and inclusion within their ranks.

2 Our approach to detect unknown biases

a. Training of the model

Our approach involved using a pretrained distilBERT model to compute attention masks and embeddings for the biographies. Subsequently, we incorporated two linear layers for job classification. This allowed us to predict job roles with a reproducible 84.5% accuracy.

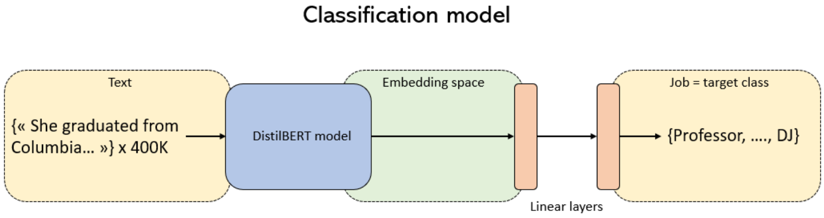

Figure 1 – Classification model architecture

b. Finding biases

Now, our aim is to scrutinize the model's errors and determine whether specific criteria (to be identified) contribute to bias. To accomplish this, we employed the test dataset from the previous training, consisting of around $N = 100,000$ English biographies. Using the distilBERT model, we obtained a set $\mathcal{B} = \{b_1, \ldots, b_N\}$ of encoded biographies and a set of attention vectors $\mathcal{A} = \{a_1, \ldots, a_n\}$. Each biography is a vector in $(\mathbb{R}^e)^s$, where $s = 512$ represents the number of tokens in each biography and $e = 768$ denotes the embedding space's size.

Our goal is to cluster these biographies and assign a discriminatory trait to each cluster. To identify potential discrimination factors, we employed GPT-3 to generate a list of possible workplace discrimination factors. The resulting list included 16 words, being ["Sex", "Ethnic", "Age", "Parent", "Divorced", "PhD", "Rich", "Health", "Religion", "Veteran", "Gender", "Race", "Old", "Married", "Poor", "Handicap"]. We aimed for short words in order to limit tokens per word, and hence the variance in the embeddings obtained post-treatment. After encoding, we obtained a set of embeddings $\{k_1, \dots k_n\}$ in $\mathbb{R}^e$.

To consolidate information for each biography, we performed a weighted average and computed :

$$\overline{b_i} = \frac{\langle b_i, a_i \rangle}{\|a_i\|_1}$$

Progress could be directly made on identifying small subsets of attention heads and applying precise attention reweighting on them, directing the model attention to user-specified parts. This approach is studied in [23] with PASTA (Post-hoc Attention STeering Approach).

For clustering, we conducted dimensionality reduction on $\overline{b_i}$ to reduce the embedding dimension from $e = 768$ to $e' = 5$, in order to avoid curse of dimensionality for upcoming distance-based clustering algorithms. Two methods were examined for this purpose : the more traditionnal PCA, and the more sophisticated t-SNE (t-distributed Stochastic Neighbor Embedding), the second one being quite so convenient to form more distinct groups.

Denoting the PCA transformation as P, we performed clustering on $\{P(\overline{b_1}), \dots, P(\overline{b_N})\}$, creating $k = 100$ clusters $C_1, ..., C_k$. For each cluster, we determined its centroid c_i and found the closest keyword in the reduced space among $P(k_1), \dots, P(k_n)$. Finally, we established a mapping α associating each cluster C_i with a keyword k_{α_i}. A decoder block was employed to retrieve the original English keyword associated with each cluster.

The overall process is summarized in the following diagram :

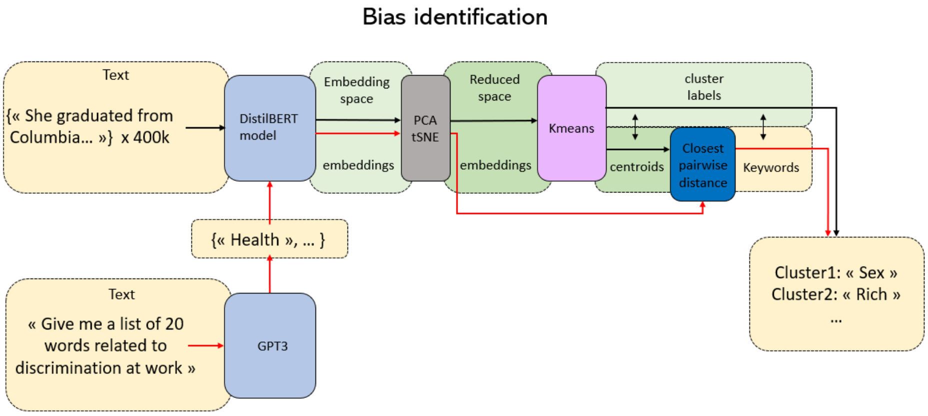

Figure 2 – Keyword association process

3 Comments and results

a. Some relevant outputs

Let's analyze the results of our clustering. We plot the true positives, false positives and false negatives for each jobs and see whether we can see some patterns related to the keywords.

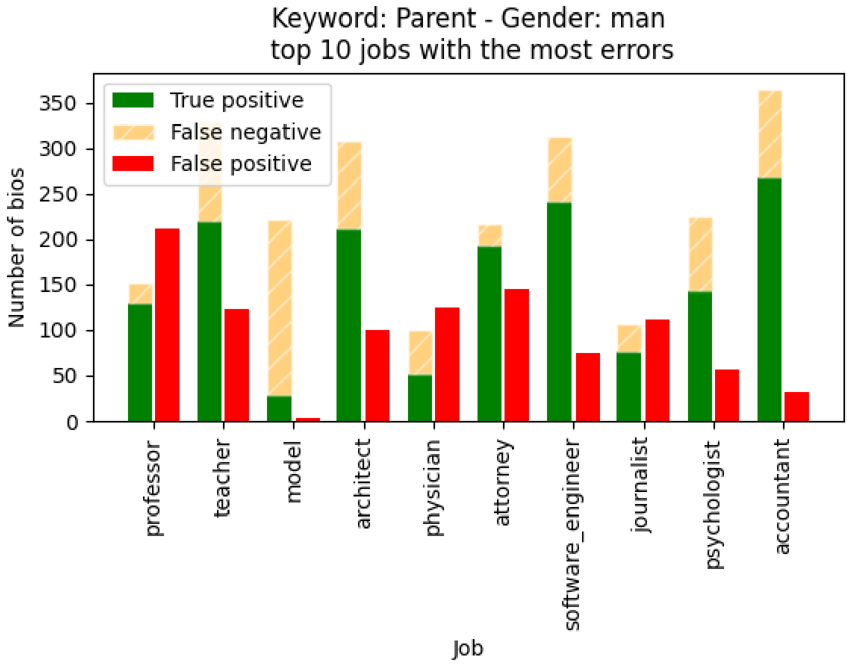
Keyword: Parent - Gender: man
top 10 jobs with the most errors

FIGURE 3 – Clusters associated to keyword "parent", with men only

Here we can see two different patterns :

• Jobs traditionally associated to women are less predicted and jobs traditionally associated to men are more predicted. For instance, there is a lot of false negative for the job "model" (traditionally associated to women) and a lot of false positive for the the job "professor" (more associated to men traditionally).

• Moreover, the biographies in this cluster are associated to the word "Parent" and therefore there are a lot of false positive for jobs associated to the idea of transmission and children (e.g. professor and teacher).

Here is a specific biography from the cluster :

> 'He's born in Brescia in ', graduated at the Brescia Conservatory and at the Istituto d'Arte of Guidizzolo, Avisco expert in A/V communication and applied arts, with particular attentions to the relation between sound, images and infant world. As he underlines in his curriculum, "he's connected to the childhood world He's convinced that searching in the A/V universe moments of free expression together with children and kids is a way to create "non-art" out of the rigid and competitive structures schemes "

FIGURE 4 – An example of biography in the cluster predicted as teacher instead of model

We can see that the concepts of childhood are give several times. This deceives the model and predict teacher instead of model. The keyword allows to find the bias induced in the dataset and maybe correct it when it's discovered.

b. Robustness of the approach

Unfortunately, while analyzing a wider range of biographies. We found out that the patterns we found were not always clear on a wide range of biographies. The fact that some patterns are emerging from this process is already interesting as such, because it was not clear at the beginning that the clusters would allow some concepts to emerge.

The main issue in our attempt is that we are comparing an averaged version of the bios (when we do the scalar product of the embeddings with the attentions) with a unique embedding associated with a concept. Those two thing are not comparable by natures, because when we average the embeddings of the sentence, the averaging can "kill" some specificities of the individual embeddings of words.

We examined the keywords and the averaged bios in a reduced space using T-SNE (similar to PCA but in dimension 2 to visualize the distances of the embedding in a small space). We obtained the following graph :

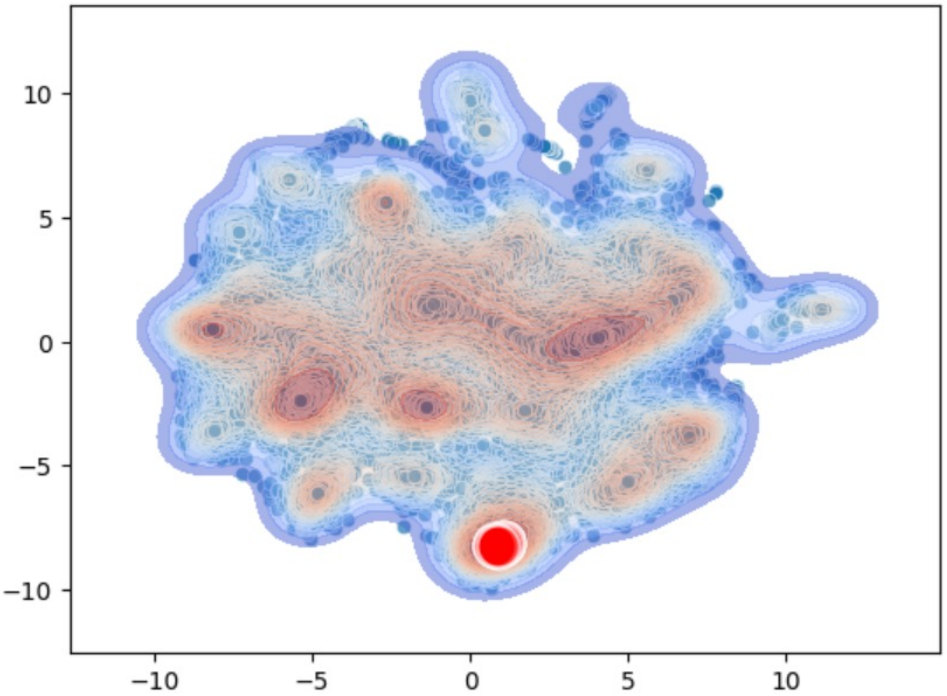

Figure 5 – Embeddings of keywords via T-SNE (red dots) and density of the bios via T-SNE

As we can see, the keywords are very close to each other whereas the averaged bios are more widespread. As we anticipated, the bios and the keyword are intrinsically different by nature and therefore are quite difficult to compare as such.

But then, why where we able to find some interesting patterns in the clusters ? Our hypothesis is that the clusters were indeed semantically close, and the relations to the keyword we observed were only a coincidence.

Finally, the clustering we did is interesting for itself because some patterns appeared, but the association with some keywords was not efficient as it is really hard to associate sentences with one concepts.
From this first attempts, we were able to derive other potential directions to find unknown biases. We discuss it this in the next section.

4 Potential improvements of our methodology and new further research directions

Despite encountering challenges in achieving reproducibility through our approach, it is essential to recognize that the insights gleaned from the obtained results can still serve as valuable assets in uncovering and addressing unknown biases. While reproducibility is a cornerstone of scientific inquiry, the absence of fully reproducible outcomes does not diminish the significance of the findings or render them obsolete. Instead, it prompts us to adopt a more nuanced perspective and explore alternative avenues for leveraging the acquired insights effectively.

Firstly, the results obtained, even if not fully reproducible, can offer valuable insights into potential biases present within the dataset or the analysis methodology itself. By critically examining the discrepancies and inconsistencies encountered, we could gain a deeper understanding of the underlying factors influencing the outcomes. These discrepancies serve as vital indicators, guiding future investigations and adjustments to enhance the robustness and reliability of the analysis framework.

While our approach may not have yielded perfectly reproducible outcomes, the insights garnered provide a foundation upon which subsequent research can build. By initiating discussions and investigations into potential biases, the results pave the way for a more comprehensive examination of the subject matter, facilitating the identification and mitigation of previously unrecognized biases. Following approaches could for instance include :

a. Normalizing averages or keywords

A first approach could be to change the way we compare the keywords to the clusters. As we seen previously, the variability of the keywords' embedding is much more reduced than the variability of the sentences' embeddings. Therefore, it could be useful to perform some normalization before comparing the keywords and the sentences embeddings. The main risk of doing this is to alter the semantic information of the embeddings. Studying different normalizations could allow to have more spread out keywords within the bios (and not concentrated at the same point as we observed with TSNE). Could be studied methods like token merging, filtering and aggregation based on previous works. The second approach might be particularly effective, considering the wide range of words completely unrelated to any potentially discriminating factor, creating a large noise.

b. Using [CLS] token

Another approach would be to use the [CLS] token, which is the first token output by BERT models from each of the bios. This token is known as the classification token, and is optimized for classification tasks on sentences. This token is supposed to summarize the information contained in one sentence via an unique token. If a semantic correspondance can be established between the [CLS] tokens and the tokens of the keywords, then there could be an opportunity to link the keywords with the bios and hence determine biases as we did previously but with the [CLS] token instead of an average.

c. Using LIME-like approach

In the context of explicability in machine learning, the "LIME" method consists in training some models on a part of the weights or altered versions of the models to locate where the information is located. Similarly, we could do the same with the words of the bios : one can alter a word of the bio (by replacing it by another word or a space for instance) and see how the performance of the model is impacted by such changes. Using this method, we could be able to determine key words in the bios that have the most importance (metric also to be discussed) in order for the classification model to predict, and then analyze these words to understand where the errors come from.

d. Using frequency analysis

Another approach would be to leverage the frequency of words appearing in bios. One common bias we observed is that some linking words can have some impact on the performance of the model. For instance, if the word "but" is used in the biography, then the first part of the biography is likely to be less related to the actual job, but rather exploring another part of the protagonist's life or occupation. Therefore, it could be useful to perform some frequency analysis on the words in the bios and see whether we can derive some correlations between the errors of the models and some recurrent words. We could also expect some of the discrimination words given by GPT such as "parent".

e. Using semi-supervised learning

Finally, if none of the above methods are applicable. It could be useful to "help" the model with some supervised learning. For instance by telling some characteristic biases presents in some bios and letting the model derive other biases. The issue with this approach is that this approach may not allow to discover biases not already given to the model and then it could be hard to determine unknown biases.

Références

[1] Matan Avitan, Ryan Cotterell, Yoav Goldberg, and Shauli Ravfogel. What changed ? converting representational interventions to natural language, 2024.

[2] Imre Z. Rusza Ben Green. Sum free sets in abelian groups. *arXiv :math/0307142v4 [math.CO]*, 2004.

[3] Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. Interactive weak supervision : Learning useful heuristics for data labeling, 2021.

[4] Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. End-to-end weak supervision, 2021.

[5] Jianfeng Chi, William Shand, Yaodong Yu, Kai-Wei Chang, Han Zhao, and Yuan Tian. Conditional supervised contrastive learning for fair text classification, 2022.

[6] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios : A case study of semantic representation bias in a high-stakes setting. 2019.

[7] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios : A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM, January 2019.

[8] Sean Eberhard. Product mixing in the alternating group. *discrete analysis journal*, 18, 2016.

[9] T. Gowers. Quasi-random groups. *arXiv :0710.3877v1 [math.CO]*, 2007.

[10] Xudong Han, Timothy Baldwin, and Trevor Cohn. Towards equal opportunity fairness through adversarial learning, 2022.

[11] Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. fairlib : A unified framework for assessing and improving classification fairness, 2022.

[12] Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. Debiasing isn't enough ! – on the effectiveness of debiasing mlms and their social biases in downstream tasks, 2022.

[13] Kiran S. Kedlaya. Large product-free subsets of finite groups. *Journal of combinatorical theory*, Series A(77) :339–343, 1997.

[14] Kiran S. Kedlaya. Product-free subsets of groups, then and now. *arXiv :0708.2295 [math.GR]*, 2007.

[15] L. Pyber L. Babai, N. Nikolov. Product growth and mixing in finite groups. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 248–257, 2008.

[16] V. T. Sos L. Babai. Sidon sets in groups and induced subgraphs of cayley graphs. *Europ. J. Combin.*, (6) :101–114, 1985.

[17] R. Rasala. On the minimal degrees of characters of s_n. *Journal of algebra*, (45) :132–181, 1977.

[18] Laurent Risser, Alberto Gonzalez Sanz, Quentin Vincenot, and Jean-Michel Loubes. Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization, 2019.

[19] Shun Shao, Yftah Ziser, and Shay Cohen. Erasure of unaligned attributes from neural representations, 2023.

[20] Shun Shao, Yftah Ziser, and Shay B. Cohen. Gold doesn't always glitter : Spectral removal of linear and nonlinear guarded attribute information, 2023.

[21] Gerard Thompson. Classifying groups of small order. *Advances in Pure Mathematics*, (6) :58–65, 2016.

[22] Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing, 2023.

[23] Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend : Post-hoc attention steering for llms, 2023.