Visual Exploration of Indirect Bias in Language Models

Judith Louis-Alexandre and Manuela Waldner
TU Wien, Austria

Abstract
Language models are trained on large text corpora that often include stereotypes. This can lead to direct or indirect bias in downstream applications. In this work, we present a method for interactive visual exploration of indirect multiclass bias learned by contextual word embeddings. We introduce a new indirect bias quantification score and present two interactive visualizations to explore interactions between multiple non-sensitive concepts (such as sports, occupations, and beverages) and sensitive attributes (such as gender or year of birth) based on this score.

CCS Concepts
- Human-centered computing → Visual analytics; Computing methodologies → Natural language processing;

1. Introduction
Natural language processing (NLP) applications such as dialogue management or machine translation are nowadays mostly based on machine learning algorithms operating on language models, such as word embeddings [MCCD13] or transformer models [DCLT19]. These language models are trained on very large text corpora, which are likely to include stereotypes. Language models learn these stereotypes in the course of their training, leading to bias in downstream applications. Indeed, it has repeatedly been demonstrated that direct (or explicit) bias in language model exists [BCZ*16, PAL20], and methods for debiasing have been proposed [BCZ*16, DLPS20]. Direct bias usually can be measured locally, for instance by varying the gender of the sentence subject (e.g., “He works as an engineer” vs. “She works as an engineer”). What is more challenging to discover is indirect bias [LWF*21]. Indirect bias is triggered by a seemingly neutral attribute, such as the Zip code, but is often caused by a correlation with some sensitive attribute [ZWW17]. Indirect bias is sometimes only evident in a holistic, global context, spanned across multiple phrases (e.g., “Aubrey is a woman. She works as a hairdresser and likes to drink tea.”) and therefore also requires correct context association [LWMS21].
The goal of our work is to enable users to interactively explore potential indirect bias in language models. We thereby followed three main design goals: discovery of indirect bias (G1) across multiple targets and attributes, exploration of multiclass bias (G2), i.e., targets and attributes beyond binary levels, such as female – male, and reasoning about potential sources of bias (G3), i.e., sensitive attributes that may explain unexpected influences of non-sensitive attributes on target variables. To fulfill our design goals, our work has the following two main contributions: (1) a new indirect bias score to probe bias beyond local single-sentence scope and (2) the design of two interactive visualizations to support visual exploration of bias between multiclass targets and sensitive or non-sensitive multiclass attributes.

2. Related Work

In their seminal work, Bolukbasi et al. [BCZ*16] showed that the vector space of word embeddings [MCCD13] can be probed for implicit bias similarly to the Implicit Association Tests (IAT) [GMS98]. A large number of bias metrics for word embeddings have afterwards been proposed, like the Word Embedding Association Test (WEAT) [CBN17] or the Relative Norm Difference [GSJZ18]. In contrast to word embeddings, contextual word embeddings like BERT [DCLT19] preserve sentence-level context and are more and more replacing traditional word embeddings [KVP*19]. Bias metrics developed for traditional word embeddings cannot consistently reproduce bias in contextual word embeddings [MWB*19, KVP*19]. A well-known bias metric for contextual word embeddings is the Sentence Encoder Association Test (SEAT) [MWB*19], which is an extension of WEAT [CBN17]. Like WEAT, it only measures binary bias between two extremes. The Logarithmic Probability bias score [KVP*19] measures the bias between a single target word and an attribute. It is more flexible than SEAT [MWB*19] because it is not operating on stereotype pairs. Therefore, our indirect bias score builds upon this score.

Bias metrics allow users to explicitly quantify an expected bias. Visualization can facilitate untargeted exploratory analysis of language models with respect to potential bias. For example, DiscrLens [WXC*21] supports interactive exploration of intersectional bias – i.e., bias caused by a superposition of several attributes – on general machine learning models. In the context of language models, researchers have visualized the association between targets, such as a set of sports or a general list of names as bridges . This way, context association [LWMS21] may compensate for the fact that some attributes are more common than others and therefore yield higher probabilities in the mask prediction task.

The direct LogProb score can only probe direct associations between target and attribute that can be expressed in a single template sentence. Our extension, the indirect LogProb bias score splits the template sentence into a pair of sentences $s_1$ and $s_2$, which are linked through a set of bridge elements $B$. For the $i$th bridge element $b_i \in B$, we construct two template sentences: One links the target to the bridge $s_1(t, b_i)$, and the other one links the bridge to the attribute $s_2(b_i, a)$. For each template sentence, the target probabilities are computed as $p_1(\text{target}) = P(\text{[MASK]} = b_i \mid s_1(t, b_i_{\text{masked}}))$ and $p_2(\text{target}) = P(\text{[MASK]} = a \mid s_2(b_i, a_{\text{masked}}))$. In addition, we compute $p_1(\text{prior})$ by masking both, bridge and target, and $p_2(\text{prior})$ by masking bridge and attribute. The bias probabilities of the sentence pair $p_1(t, b_i)$ and $p_2(b_i, a)$ are computed as log ratio of the target and prior probabilities, as for the original bias score. The final indirect LogProb score is then the Pearson correlation between the bias probabilities of all template sentence pairs generated from the bridge, as illustrated in Figure 2.

As bridge, we use a list of first names, as names naturally link targets and attributes to individuals. Names have been shown to have strong associations with valence, career, and family [CBN17]. Intuitively, the indirect LogProb bias score is high if the predicted probability of bridge names is high for target and attribute – or low for both. As a representative bridge for the U.S., we selected a list of the 100 most frequent female and male baby names given in the U.S. between 1920 and 2020 [Soc22] with 779 names in total.

A strength of the indirect LogProb bias score is that the template sentences can be very short because target and attribute queries are expressed in two independent sentences – especially when using names as bridges. This way, context association [LWMS21] may be preserved. Furthermore, it can measure bias even for targets that are unknown to the model’s vocabulary (such as fireman or police officer in case of BERT) and therefore will not be predicted by the model in a mask prediction task like “Jim works as a [MASK]”.

To validate the indirect LogProb bias score, we compare its predictions to those of the direct LogProb [KVP*19] score. We tested the predictions on a BERT model [DCLT19], pre-trained on the Wikipedia dump dataset and the BookCorpus dataset [ZKZ*15]. Since the LogProb score is sensitive to the formulation of the tem-
plate sentence, we always constructed multiple variants of the two sentences $s_1$ and $s_2$ and averaged their target and prior probabilities.

First, we compare the direct and indirect LogProb bias scores by predicting beverage preference by gender and comparing the predictions to a public health study [BCCW’16]. We therefore link the four alcohol beverages discussed in the public health study (beer, wine, liquor, and alcohol in general) to the gender-defining word woman and compute both bias scores. According to the ground truth [BCCW’16], only for wine there is a higher consumption ratio for women than for men. The indirect score shows the same trend, while the direct score predicts a positive bias towards woman for all four alcoholic beverages. It is also notable that, out of a list of 18 popular alcoholic and non-alcoholic beverages, the direct LogProb bias score predicts the same top-two beverages for males and females (namely champagne and whiskey). The indirect score predicts tea and milk as the beverages most strongly associated with women, but beer and liquor for men.

Second, we test the association between beverage preferences and occupations, and whether there might be an indirect gender bias. We test 99 occupations (inspired by prior work [BCZ’16, LMW’20]) as targets × 10 beverages as attributes. For the indirect score, we can observe that milk and tea are associated with occupations like nanny and businesswoman, which also positively correlate with woman. Similarly, beer and liquor are associated with occupations like fireman or mechanic, which also correlate with man. This indicates that the sensitive attribute gender could indeed explain the non-sensitive association between occupations and preferred beverages revealed by the indirect score. The direct score, on the other hand, predicts milk to be the most preferred beverage by farmer. Similarly, it finds an association between fireman and water. One potential explanation is that, due to the concatenation of two phrases into one sentence, the context association [LWMS21] (in this case the fact that we want to investigate what people like to drink rather than what they work with) might get lost.

### 4. Visual Bias Exploration Interfaces

The bias data we are dealing with can be represented as multidimensional tabular data, where each cell represents the indirect LogProb bias score of a target-attribute combination. We therefore experimented with two simple visualizations that are commonly employed for such data characteristics: (1) a table view and (2) a scatterplot. In our prototype, we support the following set of non-sensitive concepts: 99 occupations, 18 beverages, 30 sports, 10 countries, 618 mental and physical traits (extended from [KVP’19]). In addition, we included sensitive attributes like gender, country of origin, race, or age.

**Table-based visualizations** are an intuitive choice for multidimensional data structures as they essentially represent a 2D projection of the higher-dimensional data cube [STH02]. Each cell can serve as nested display or show a single associated value as text label or color. The indirect LogProb bias score associated with the respective target-attribute combination is shown through a diverging color map (Figure 1). By clicking on the cell, the associated scatterplot visualizing the direct bias scores for target and attribute across all bridge elements is shown (see Figure 2 right). Users can choose any target-attribute combination from a dropdown menu so that they can explore potential indirect bias (G1).

However, the table is limited by the number of items that can be effectively shown on the screen – especially if rows and columns need to be labeled, as in our case (violating G2). We solve this through interactivity: initially, the table shows all target and attribute levels in a pre-defined order (e.g., alphabetically). The user can then select a target to sort the attribute rows according to the bias score with the selected target item. This limits the number of displayed rows to the five most positively and negatively correlated attributes (Figure 1(A)). This allows the users to express queries like “Which traits are associated with homemaker, and which are not?” By clicking the selected column header a second time, the system also sorts the columns based on the cosine similarity of the targets’ attribute vectors to the selected target. This supports questions like “Which professions are supposedly done by people with opposite characteristics to homemaker?”, as shown in Figure 3.

Table sorting remains persistent when changing the displayed attributes – even if the attribute based on which the table is sorted is no longer visible. This way, users can visually test whether a correlation between a target and a non-sensitive attribute is potentially caused by a sensitive attribute (G3). In other words, they can perform visual queries like “Are occupations that are considered to be done by ambitious people also predominantly done by males?”

**Scatterplots** are popular for inspecting clusters in high-
and required some explanation by the study conductor. 

indirectly explain discovered bias was considered a difficult task 

For both visualizations, reasoning whether a sensitive attribute can 
ties to find a specific target, which required hovering over the dots. 

the users during their investigation. Users also mentioned difficul-

always immediately understood and was considered rather little by 

sort the table. For the scatterplot, the 2D layout of the dots was not 

expression of the of both visualizations was rated as positive by the 

Figure 3: Table view: traits positively correlated with homemaker (rows) as well as the most similar and dissimilar occupations to homemaker (columns) based on their associated traits.

dimensional data. Using dimensionality reduction algorithms, the 

multi-dimensional tabular data is reduced to two spatial dimensions 
to show the similarity of data items in 2D [SMT13]. In our design, 
each scatterplot dot represents a single target item, such as an occupa-
pation, a beverage, or a type of sport. The high-dimensional feature 

vector represents the bias scores associated with a single target and 

the potentially large number of attribute levels (G2), e.g., the indir-

ect LogProb bias scores for sports with respect to the 618 mental 

and physical traits associated with the people practising them (see 

Figure 1(B)). This way, users can inspect similarities of targets with 

respect to the selected attribute based on their proximities in the 
scatterplot. This allows to answer questions like “Which occupa-
tions are similar with respect to the sports these professionals like 
to do?” (see Figure 4). Compared to the table view, the scatterplot 
scales much better with the number of target items to be shown. The 

target labels can be revealed as tooltips by hovering the scatterplot 
dots with the mouse (see Figure 1(B) top). Like for the table view, 

any target-attribute combination can be chosen to explore indirect 

bias (G1).

However, the scatterplot alone cannot answer questions like “In which occupations do people like to play baseball in their free-time?” To support such queries, we allow users to select any sensi-
tive or non-sensitive attribute level to define the colors of the scat-
terplot dots (Figure 4). If the color changes systematically with the 
dots’ positions, we can assume that the selected attribute level can 
explain the similarities between the target items. This way, a single 

sensitive attribute value (e.g., female) can be directly probed on 
a visualized target-attribute scatterplot (G3). This supports queries 
like “Are occupations that are associated with similar sports also 
predominantly associated with males / females?”

5. Preliminary Results and Conclusions

We conducted a preliminary qualitative study with ten volunteers 
(five females, aged 22-25), wherein five explored potential bias us-

ing the table view and five with the scatterplot. The overall im-

pression of the of both visualizations was rated as positive by the 

users. User feedback indicates that the table view was easier to un-
derstand initially, but some users would have liked more options to 
sort the table. For the scatterplot, the 2D layout of the dots was not 
always immediately understood and was considered rather little by 
the users during their investigation. Users also mentioned difficul-
ties to find a specific target, which required hovering over the dots. 
For both visualizations, reasoning whether a sensitive attribute can 
indirectly explain discovered bias was considered a difficult task 
and required some explanation by the study conductor.

Direct bias was easier to understand. Some of the users’ prior 
expectations with respect to bias could be confirmed, for instance 
an association between artist and passionate. Other associ-
ations were unexpected, yet seemed plausible for the users, such as 
a preference for champagne and France as country of origin. 
However, some associations were questioned by our users, such as 
engineer with homosexual and the top countries associated 
with beer (Vietnam and the United States).

In summary, while the indirect LogProb bias score could reliably 
detect direct and indirect bias in our quantitative experiments, it 
occasionally delivered questionable associations in the exploratory 
study. Intuitively, given names may not be able to predict some 
attributes, like sexual orientation. Also, the bridge currently has a 
strong focus on popular names in the U.S., and therefore may fail 
on sensitive attribute predictions like country of origin or race. In 
the future, we will therefore investigate more international names 
or even completely alternative bridge sets.

We further showed how to enrich two common visualizations 
with interactivity to support our design goals. Our preliminary 
study shows that both visualizations are suitable for our intended 
tasks, but the table view is probably easier to understand initially. 
In the future, interaction with the scatterplot view could be facili-
tated by ranking attributes used for color-coding based on quality 
metrics [SA15].

Acknowledgements

This work is partially supported by the Austrian Science Fund 
(FWF): P 36453.