WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings

Bhavya Ghai
Stony Brook University
bghai@cs.stonybrook.edu

Md Naimul Hoque
Stony Brook University
mdhoque@cs.stonybrook.edu

Klaus Mueller
Stony Brook University
mueller@cs.stonybrook.edu

ABSTRACT

Intersectional bias is a bias caused by an overlap of multiple social factors like gender, sexuality, race, disability, religion, etc. A recent study has shown that word embedding models can be laden with biases against intersectional groups like African American females, etc. The first step towards tackling such intersectional biases is to identify them. However, discovering biases against different intersectional groups remains a challenging task. In this work, we present *WordBias*, an interactive visual tool designed to explore biases against intersectional groups encoded in static word embeddings. Given a pretrained static word embedding, WordBias computes the association of each word along different groups based on race, age, etc. and then visualizes them using a novel interactive interface. Using a case study, we demonstrate how WordBias can help uncover biases against intersectional groups like Black Muslim Males, Poor Females, etc. encoded in word embedding. In addition, we also evaluate our tool using qualitative feedback from expert interviews.

CCS CONCEPTS

• **Human-centered computing** → **Visual analytics**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Algorithmic Fairness, Visual Analytics, Word Embeddings

1 INTRODUCTION

Word embedding models such as Glove [33] and Word2vec [30] can be understood as a mapping between a word and its corresponding vector representation. They serve as the foundational unit for many NLP applications such as sentiment analysis, machine translation, etc. and could possibly be used to bootstrap any NLP task [3]. It has been shown that word embedding can learn and exhibit social biases based on race, gender, ethnicity, etc. that are encoded in the training dataset [4, 7, 17]. Social biases in word embeddings are manifested as stereotypes or undesirable associations between words [17]. For example, word embedding models might disproportionately associate Male names with career and math, while Female names might be associated with family and arts [17]. Existing literature has mostly focused on measuring and mitigating the *individual* social biases based on race, gender, etc. encoded in word embeddings [4, 7, 14, 17, 20, 26, 42].

Recent studies have shown the presence of *Intersectional Bias* in AI systems [5, 20, 24] i.e. a bias towards a population defined by multiple sensitive attributes like 'black muslim females' [9, 12]. For example, facial recognition software applications have been shown to perform worse for the intersectional group 'darker females' than for either darker individuals or females [5]. Similarly, word embedding models have also been shown to contain biases against intersectional groups like Mexican American females [20]. When such biased word embeddings are used for any downstream application, their inherent social biases are propagated further, which can cause discrimination [32, 49]. Hence, it becomes critical to investigate the presence of different intersectional biases before using it for some application.

Stereotypes associated to an intersectional group say 'Black Males' are composed of stereotypes pertaining to constituting subgroups (Blacks and Males) along with some unique elements [18]. The proportion of stereotypes which overlap with either of the constituting subgroups can vary based on the intersectional group. For example, a study on 627 undergraduate students found that the percentage of overlap for intersectional groups like White men is 81%, White women is 88%, Black women is 44%, Black men is 73%, Middle Eastern American men is 91%, etc. [18]. This work focuses on this overlapping aspect of intersectionality. Given that word embedding models can consist of thousands of unique words and the number of intersectional groups can increase drastically with the number of sensitive attributes considered, it becomes challenging to explore the massive space of possible associations. Writing custom code to test the different associations can be tedious and ineffective.

In this work, we present the first interactive visual tool, *WordBias*, for exploring biases against different intersectional groups encoded in word embeddings. Given a pretrained word embedding, our tool computes the association (bias score) of each word along different social categorizations (bias types) like gender, religion, etc. and then visualizes them using a novel interactive interface (see Figure 1). Here, each categorization (bias type) e.g. race consists of two subgroups, say Blacks and Whites. Using bias metrics, WordBias computes the degree to which a word aligns with one subgroup over the other. The visual interface then allows the user to *investigate* how a specific word associates with different individual subgroups and also *discover* words that are associated with an intersectional group. Considering the overlapping aspect of intersectionality, WordBias considers a word to be associated with an intersectional group say 'Christian Males' if it associates strongly with each of its constituting subgroups (Christians and Males).

Users can interact with our tool to explore the space of word associations and then use their real world knowledge to determine if a given association is socially desirable. For example, the association between the word 'queen' and female is desirable whereas the association between 'teacher' and female is not. Using a case

Conference'17, July 2017, Washington, DC, USA
2021. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

study, we demonstrate how WordBias can help discover biases against different intersectional groups like 'Young Poor Blacks', 'Black Muslim Males', etc. in Word2Vec embedding. Identifying such biases can serve as the first step toward deterring its spread and help develop counter-strategies. Lastly, we evaluate the usability and utility of our tool using qualitative feedback from domain experts. We have made the source code for our tool along with a live demo publicly available for easy reproducibility and accessibility (github.com/bhavyaghai/WordBias).

2 RELATED WORK

2.1 Bias in Word Embeddings

The existing literature on bias in word embeddings can be broadly classified into bias identification and mitigation. For bias identification, a number of bias metrics are proposed like *Subspace Projection* [4], *Relative Norm Difference* [17], *Word Embedding Association Test (WEAT)* [7], etc., but there is no single agreed-upon method [48]. Our tool builds upon such bias identification metrics to explore the space of word associations and help detect biased associations. More specifically, our tool uses the *Relative Norm Difference* metric as it is simple to interpret and can be easily extended for different kinds of biases. Previously, this metric has been used to capture biases against individual sensitive groups like females. In this work, we have used this metric to capture biases against intersectional groups as well. Once bias has been detected, there are a host of debiasing techniques which can be used for bias mitigation [4, 43, 49]. However, we will not go into these details as our work is limited to bias discovery. Our work relates more closely with Swinger et al. [40], who tries to find biases in word embeddings using purely algorithmic means compared to our visual analytics approach. Our dynamic visual interface makes the entire process more interactive and accessible to non-programmers. It also provides more flexibility by allowing the user to drive the bias discovery process as they see fit.

2.2 Visual Tools

Recent years have seen a spike in visual tools aimed at tackling Algorithmic fairness like Silva [47], FairVis [6], FairSight [1], What-If [45], etc. All of these tools help detect Algorithmic Bias but they are mostly limited to tabular datasets. Moreover, many of these tools are designed to deal with individual biases and not intersectional biases. Our tool, WordBias, helps fill in this gap by helping discover intersectional biases encoded in word embeddings. Our tool relates closely to Google's Embedding Projector (GEP) [38] which supports a custom projection adopted from [4] to visualize bias. As a general-purpose tool primarily aimed at visualizing high dimensional data in 2D or 3D space, GEP has several limitations when it comes to exploring biases in word embeddings: (1) it does not support any bias quantification algorithm, (2) it is limited to visualizing only two types of bias simultaneously, and (3) its custom projection only allows one word to characterize a subgroup, say 'he' for males. In this work, we have tried to overcome all these limitations by carefully designing an interactive visual platform geared towards exploring social biases.

3 WORDBIAS

3.1 Design Goals

Based on the current literature and the problem at hand, we have identified the following four design goals:

- **G1. Bias Scores**: Our tool should compute bias scores and accurately visualize them such that the user can quickly identify the different subgroups a word is associated to along with their degree of association.
- **G2. Bias Exploration**: Our tool should support quick and intuitive exploration of words associated with a single subgroup say Males or an intersectional group say *Rich White Females*.
- **G3. Bias Types**: The existing literature on biases in word embeddings is heavily skewed towards gender bias (93%) followed by racial bias (54%) [37]. Our tool should support the exploration of these well known biases but also under-reported biases based on physical appearance, political leanings, etc. or any user-defined bias type.
- **G4. Data Volume**: Word embedding models can consist of millions of unique words. Our tool should be designed to deal with a large volume of data at both the back and the front end to ensure a smooth user experience.

3.2 Bias Quantification

We have used the *Relative Norm Difference* [17] to quantify the association (bias) of a word along different bias types. Like most bias metrics, it assumes that a given bias type, say gender, consists of two subgroups, say males and females. Each of such subgroups is defined using a set of words called *group words*. For example, group words for males might include he, him, etc. while for females, it might include she, her, etc. Mathematically, a subgroup is expressed as the average of word embeddings for the words which define that subgroup. For a given bias type, let $\vec{g1}, \vec{g2}$ represent either subgroups. We then define the bias score for a word w with embedding $\vec{w}$ as follows:

$$Bias\_score(w) = cosine\_distance(\vec{w}, \vec{g1}) - cosine\_distance(\vec{w}, \vec{g2}) \quad (1)$$

A bias score can be understood as the association of a word toward a subgroup with respect to the other. The magnitude of the bias score represents the strength of the association and the sign indicates which subgroup it is associated to. We compute bias scores for each word across bias types using Equation 1 and then repeat this process for all words.

3.3 Feature Scaling

Visualizing raw bias scores might be difficult to interpret and compare because the distribution of bias scores varies across bias types. For example, a 0.3 bias score for gender bias might mean a much stronger/weaker degree of association compared to the same score for race bias. To cope, WordBias supports two kinds of feature scaling methods namely, *Min-Max Normalization* and *Percentile Ranking*. Min-Max Normalization ensures that bias scores across bias types share the same range by simply stretching raw bias scores over the range [-1,1]. However, it is still difficult to compare bias scores because of the different standard deviation across bias types. To overcome this limitation, we use *Percentile Ranking*. Each

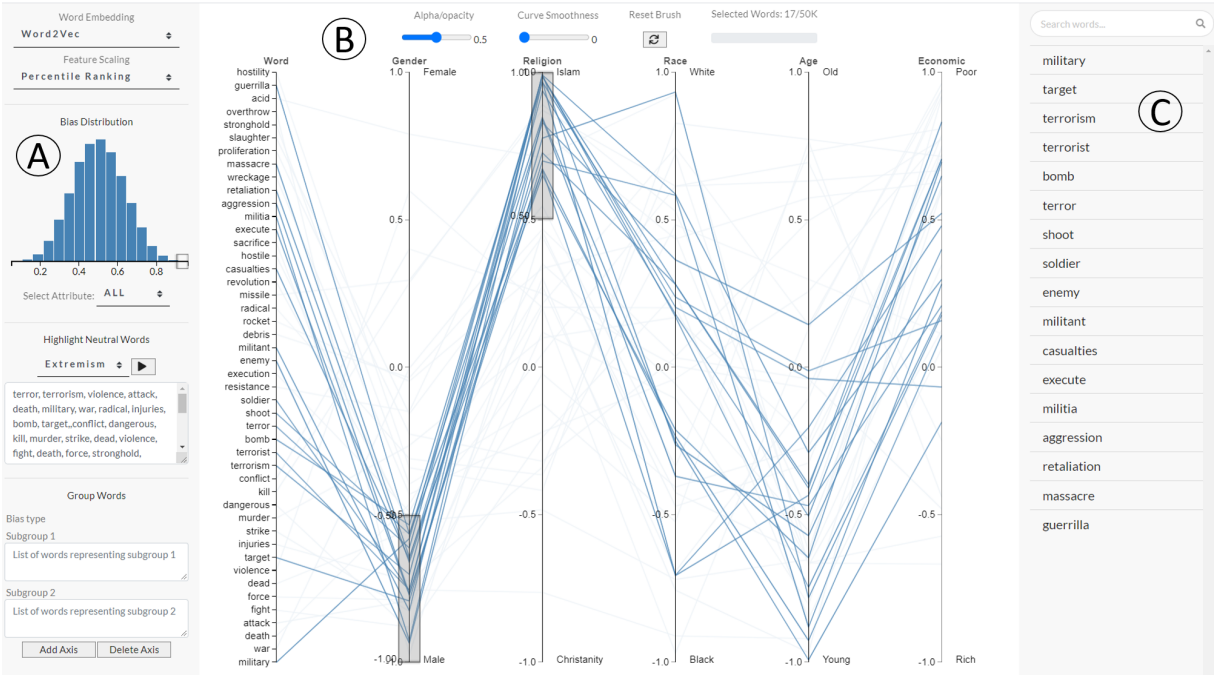


Figure 1: Visual interface of *WordBias* using Word2Vec embedding. (A) The Control Panel provides options to select words to be projected on the parallel coordinates plot (B) The Main View shows the bias scores of selected words (polylines) along different bias types (axes) (C) The Search Panel enables users to search for a word and display the search/brushing results. In the above figure, the user has brushed over 'Male' and 'Islam' subgroups. Words with strong association to both these subgroups are listed below the search box.

word is assigned a percentile score [46] based on its ranking within its subgroup. For e.g., a 0.8 percentile score means that 80% of all words associated with the same subgroup have a bias score less than or equal to the given word. This makes it easier to interpret and compare bias scores across different bias types (**G1, G2**). It should be noted that percentile scores can sometimes be misleading as they are not equally spaced. Lets say that raw bias scores for most words along a bias type is close to 0. However, we can still obtain high percentile scores for words which otherwise have negligible raw bias scores. Hence, we recommend trying both feature scaling methods to get a comprehensive picture.

3.4 Design Rationale

The problem of visualizing biases against intersectional groups boils down to visualizing a large multivariate dataset where each word corresponds to a row and each column corresponds to a bias type. A straightforward solution for visualizing such high-dimensional data is to use standard *dimensionality reduction* techniques like MDS, TSNE, biplot, etc. and then use popular visualization techniques like scatter plot. However, Algorithmic bias is a sensitive domain; we must make sure that we *accurately* depict the biases of each word (**G1**). Hence, *dimensionality reduction* and related techniques like the Data Context Map[8] are not an option because they almost always involve some information loss. Using such techniques might inflate/deflate real bias scores which might mislead the user.

Next, we enumerated other possible ways to visualize multivariate dataset, like scatterplot matrix, radar chart, etc. and then started filtering these options based on the design challenges G1-G4. The scatter plot is a popular choice which is also used in Google's Embedding projector [38], but it is limited to three dimensions. A couple of more dimensions can be added by encoding radius and color of each dot yielding a plot that can visualize 5 dimensions; but such a plot will be virtually indecipherable. The scatterplot matrix can also be an option but it is more geared to visualizing binary relationships than the feature value of each point. Moreover, it becomes more space inefficient as the number of dimensions grow. Another alternative can be the biplot but it can be difficult to read and involves information loss. The radar plot provides for a succinct representation to visualize multivariate data but it can only handle a few points before polygons overlap and it becomes unreadable (defeating G4). We ended up with the parallel coordinate (PC) plot [21] based on our design goals **G1-G4**. PC can visualize a significant number of points with multiple dimensions without any information loss (G1, G4). It also facilitates bias exploration and adding new bias types. To support plotting large numbers of points, we chose canvas over SVG and also used progressive rendering [22] (**G4**).

3.5 Visual Interface

The visual interface can be classified into 3 components (see Figure 1). Next, we will discuss each component in detail.

3.5.1 Main View. At the very center is the Main View which consists of a parallel coordinate plot [21] (see Figure 1 (B)). Each axis of

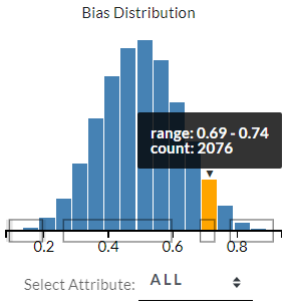


Figure 2: Select words based on their bias scores by brushing on the x-axis of the histogram.

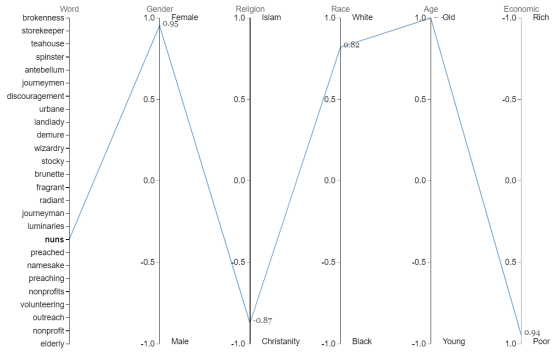


Figure 3: On Hovering over the word 'nuns', we can observe its association with 'Female', 'Christianity', 'White', 'Old' and 'Poor' subgroups.

the PC plot represents a type of bias based on gender, race, etc. and each piecewise linear curve, called *polyline*, encodes a word. Either end of each axis represents a subgroup. For example, the gender axis encodes males and females on either extremes. The higher the magnitude of a word's bias score, the higher is the inclination of the corresponding polyline towards either group. We also have an additional axis, *word*, which lists all the words currently displayed. On hovering over any word on the *word* axis, its corresponding polyline gets highlighted (see Figure 3). This visualizes all different associations for the specific word (**G1**). On clicking over any word on the *word* axis, the word and its synonyms get highlighted. Synonyms for a word are fetched via Thesaurus.com (using an API call) and from the nearest neighbors in the word embedding space.

To identify words with strong association toward a subgroup say females, the user can simply brush on the corresponding end of a given axis. Similarly, brushing either ends on multiple axes will help discover words associated to that intersectional group (**G2**). As shown in Figure 1, the user may brush on *Male* and *Islam* ends on gender and religion axes to obtain words related to the intersectional group *Muslim males*. Words selected via brushing are displayed under the search box (see Figure 1 (C)). At the top of the Main panel, there are sliders to customize *Alpha/Opacity* and *Curve smoothness* of the polylines. They are useful to see the underlying pattern between lots of polylines and to deal with the 'crossing problem'[19] respectively.

3.5.2 Control Panel. The left panel (see Figure 1 (A)), the Control Panel, allows the user to control what is displayed on the parallel coordinates plot. The user can choose the word embedding and the feature scaling method from the respective dropdown menus. It also contains a histogram accompanied by a dropdown menu. The dropdown menu contains a list of all bias types currently displayed in the parallel coordinates along with an *ALL* option. The histogram serves two purposes. First, it helps users understand the underlying distribution of bias scores for the selected bias type across the word embedding. Second, it helps users deal with the problem of over-plotting by acting as a filtering mechanism (**G4**). The user can select single/multiple ranges of variable length on the x-axis of the histogram (see Figure 2). The words whose bias score falls in the selected range(s) are displayed on the parallel coordinates. The *ALL* option (default) paints an aggregate picture as it corresponds to the mean absolute bias score across all bias types.

We precomputed commonly known biases based on gender, race, religion, age, etc. to jump-start the bias discovery process when the tool first loads (**G4**). We used group words from the existing literature for each bias type [7, 17, 25] (see Appendix). The user is free to investigate a new bias type or drop an existing one by using the Add/Delete axis button (**G3**). To add a new bias type say *political orientation*, the user needs to fill in details like axis name, subgroup names and *group words* under the 'Group Words' section and click 'Add Axis'. Here, the *Group words* should be chosen carefully as they play a critical role in computing the bias scores.

Lastly, we included a set of neutral words corresponding to categories like professions, personality traits, etc. which should ideally have no association with bias types like gender, race, etc. These words have been derived from existing literature [7, 17] (see Appendix). On clicking the 'play' button, the currently selected set of neutral words will be highlighted in the PC plot. This provides a quick way to audit an embedding for potential biases.

3.5.3 Search Panel. The right panel (see Figure 1 (C)) enables a user to search for a specific word and see the respective search/brushing results. A user can simply lookup how a specific word associates with different groups by searching for it in the search box (**G1**). This will highlight the specific word and its synonyms in the parallel coordinates. The area under the search box is used to populate the list of synonyms and brushing results.

4 IMPLEMENTATION

WordBias is implemented as a web application built over python based web framework *Flask*. On the back end, we used *gensim* package to deal with word embeddings and *PyThesaurus*[1] to fetch synonyms from Thesaurus.com. For the front end, we used D3 based library *Parallel Coordinates*[2] and used *D3.js*, *Bootstrap*, *noUiSlider*, etc. for rendering different visual components.

5 CASE STUDY

Let us assume a user, Divya (she/her), who works as a Data Scientist for a big Tech firm. Her team is tasked with building an

[1] pypi.org/project/py-thesaurus
[2] github.com/syntagmatic/parallel-coordinates

Table 1: Words with strong association with each intersectional group (within top 25 percentile of each constituting subgroup) in Word2vec embedding trained over Google News corpus.

Intersectional Group	Associated Words
Poor - Young - Black	disaster, struggle, tackle, chaos, woes, hunger, uprising, desperation, insecurity, rampage, roadblocks, scarcity, calamity, homophobia, shoddy, falter, jailbreak, mineworkers, marginalization, evictions
Rich - Old - White	formal, attractive, appealing, desirable, castle, desserts, seaside, golfing, cordial, bungalow, fanciful, warmly, salty, nutty, gentler, aristocratic, snug, prim, urbane
Black - Muslim - Male	gun, assassination, bullets, bribes, thugs, looted, dictators, electrocuted, cowards, agitating, storekeeper, looter, bleeping, lynch, strongman, disbelievers, hoodlums
Young - Christian - Male	career, dominant, brilliant, lone, terrific, heroes, superb, epic, monster, prowess, heavyweights, excelled, superstars, supremacy, fearless, inexperience, mastery, crafty, ply, conquering, rampaging
Poor - Female	ostracism, brokenness, mortgages, eviction, brothels, witchcraft, traumatized, discrimination, layoffs, uninsured, sterilizations, abortion, powerlessness, sufferer, neediest, prostitution, microloans, distressed, homelessness, miscarry
White - Christian - Female	romantic, nuns, virgin, republicans, peachy, platonic, convent, radiant, unspoiled, unpersuasive, soppy, honeymooning, drippy, soapy

automatic language translation tool. Made aware by the infamous *Google Translate* example [35, 39], she knows that such translation tools can be discriminatory toward minorities and can pose serious challenges for her organization. One of the ways in which bias can creep in is via word embeddings [4, 13]. So, she needs to audit the word embedding for different social biases before using it. One way to explore/detect biases can be via purely algorithmic means i.e., writing custom program to test the different associations. Given that exploration is a dynamic process, so one might need to tweak and re-run the code repeatedly which can be tedious and cause delays. Moreover, analysing raw numbers for thousands of words across multiple bias types can be overwhelming and ineffective. Interactive visualization techniques excel at exploratory data analysis as they provide a faster, efficient and user friendly way to interact with massive datasets effectively [23]. So, Divya decides to use a visual analytics based tool, *WordBias*, to audit her word embedding. Note that while we have used an embedding generated by word2vec [30] trained over the Google News corpus, in a real world scenario this may be a word embedding trained over the company's private data.

On first loading the tool, Divya observes that a small fraction of words are visualized which have strong association with multiple groups. These words correspond to the right tail of the histogram, i.e. they are words with high mean bias score. She hovers over some words like storekeeper, landlady, luminaries, nuns, etc. on the word axis to see their corresponding associations. Some of the associations are accurate and align well with her real world knowledge, like 'landlady' and 'nuns' have a strong association with females. In contrast, other associations, like 'storekeeper' and 'luminaries' have a strong male orientation which she views as problematic. It indicates that this word embedding might encode gender bias. To make sure that it is not a one-off case, she searches for the word 'corrupt' in the Search panel. Just by looking at the parallel coordinates display, she can make out that the word 'corrupt' and most of its synonyms like corruption, corrupted, crooked, unscrupulous, etc. have a strong association with Males and Blacks. This reaffirms the presence of gender bias and also indicates racial bias and bias against Black Males.

She carries on her investigation using different sets of words under the 'Neutral Words' section in the control panel. Each time she finds a strong association of 'ideally neutral' words with at least one kind of subgroup. When visualizing a set of *Professions*, she finds words like teacher, nurse, dancer, etc. on brushing over the female subgroup and words like farmer, mechanic, physicist, laborer, etc. on brushing over the male subgroup. Figure 1 represents the case when she chooses to visualize words characterizing *Extremism*. On brushing over the Male and Islam subgroup, she observes words like terrorist, bomb, aggression, etc. in the search panel. After this exercise, she is confirmed that this embedding encodes strong social biases against different groups as well as intersectional groups like Black males, Muslim males, etc. Her team might have to use different debiasing techniques before actually using this word embedding.

The first step towards debiasing a word embedding is to identify the different impacted groups [4, 28]. So, she explores different intersectional biases by selecting all the words using the histogram and then brushing over different subgroups. She finds lots of positive and negative stereotypes (biases) against multiple intersectional groups. Some of the more striking associations are described in Table 1. Overall, our tool helped Divya and her team to prevent a possible disaster by making them aware about the different social biases encoded in the word embedding. From here on, they can take multiple paths like trying to mitigate these biases, using a different word embedding, etc. They also need to be cautious about other possible sources of bias [29] like training dataset to make sure that bias does not creep in.

6 EXPERT EVALUATION

We conducted a set of individual 45-60 min long semi-structured interviews with five domain experts. All experts E1-E5 are faculty members affiliated to departments like Computer Science (E1, E3), Sociology (E4, E5) and Business School (E2) at reputed R1 Universities. They have taught course(s) and/or published research paper(s) dealing with Algorithmic Fairness/Intersectionality. Each expert

was briefed about the problem statement and existing solutions. Thereafter, we demonstrated the different features, interactions and the workflow of our system using a case study. Lastly, we solicited their comments on usability, utility, and scope for future improvements which are summarized as follows.

All experts found the interface to be *intuitive* and *easy to use*. Some experts found the interface to be a bit 'overwhelming' at first glance. They were unsure of where to start interacting with the tool. However, a brief tutorial neutralized these concerns. E4 commented, *"Once you understand the tool, its very useful and you know what you are seeing"*. E3 commented that *the UI looks clean* and *actions required to accomplish tasks are simple and straightforward*. E2 commented, *"Given a brief tutorial, most people should be able to get along quickly"*.

On the utility front, E2 and E3 found this tool *"Definitely useful"* for the NLP community while E1 stressed its utility for the Socio-Linguists and as an educational tool. E3 emphasized its *broad* utility for developers, researchers and consumers, and also expressed interest in using this tool for teaching about bias in their NLP class. E4 emphasized the tool's utility for researchers and showed interest in loading their own custom word embedding into the tool. E4 added, *"Anytime we want to ask a question from the data, we need to rerun the jupyter notebook which might take some time. This tool can cut down that Long feedback loop while providing rich information"*. E2 and E4 particularly liked that with WordBias users can dynamically add a new bias type on the go. This would make WordBias capable of supporting *sentiment analysis* by encoding positive and negative sentiments on either extremes of an axis. Another important aspect of Wordbias which received appreciation is its *accessibility* i.e., our tool can be hosted on a web server and then be easily accessed via a web browser without needing to install any software or dealing with github.

For the future, most experts suggested to extend support for Contextualized word embeddings like BERT [15], ELMo [34], etc. They pointed out that WordBias' current setup assumes a binary view of the real world since it only supports two subgroups per bias type. However, the real world is multi-polar. They suggested to accommodate multiple subgroups like Whites, Blacks, Hispanics, Asians, etc. under a single bias type, say race. E1 highlighted that some of the bias variables like race and economic status might be correlated. Future work should account for such correlations while computing the bias scores. E5 suggested to encode multiple word embeddings representing different time periods on each axes. This will help in analysing how different biases evolve over time. E4 suggested to add a 'Download' button which can help store all words currently displayed in the tool along with their bias scores in CSV format.

7 DISCUSSION, LIMITATIONS & FUTURE WORK

Scalability. We will discuss scalability on two aspects i.e. front-end rendering and back-end computation. On the frond end, we have used the parallel coordinate plot which can get cluttered as the number of points increases beyond a threshold. We have used a number of visual analytics based techniques to ameliorate this issue, such as histogram based selection, changing opacity of lines, brushing, highlighting words on hover, etc. We have also used canvas based progressive rendering instead of SVG to render large data effectively (G4). Finally, there are also natural limitations on the number of bias types (axes) that can be differentiated in terms of their word associations.

Our current back-end can deal with words on a scale of 10^4 while still maintaining a smooth user experience. As the number of words increases, the time for loading the word embedding and the time to calculate bias scores for a new bias type increases proportionally. Future work might use databases to store and query word embeddings to reduce load time. Furthermore, leveraging multiple compute cores will enable faster computation for any new bias type on the fly.

Quantifying Bias. Measuring bias in word embeddings is an active research area and there is no consensus on a single best metric. In our case, we have used the *Relative Norm Difference* metric. So, the bias scores reported by our tool are susceptible to the possible limitations of this metric and the group words used. The feature scaling methods, especially percentile ranking, can impact the perceived strength of an association. We recommend switching between different feature scaling methods (including raw bias scores) to get an accurate picture. Moreover, WordBias assumes a binary view of an inherently multi-polar world. This can impact the bias scores of words which do not fit into either categories. For eg., our tool reports white (race) orientation for the word 'asian' even though its a different race altogether. One must interpret the bias scores responsibly in light of these limitations. Future work might support multiple bias metrics to paint a more comprehensive picture and also include metrics which can better capture the multi-polar world.

It is important to understand that the the term *Intersectionality* has a broader meaning beyond multiplicity of identities [10, 11, 16]. Quantifying such a complex sociological concept accurately needs more research. Our tool considers a narrow definition of Intersectionality where a word is linked to an intersectional group only if its relates strongly with each of the constituting subgroups. In reality, there can be cases like 'Hair Weaves' where a word is associated with an intersectional group (Black Females) even though it does not relate strongly with either constituting subgroups (Blacks or Females) [18]. Future work might incorporate bias metrics like EIBD [20] which can capture such cases as well.

Utility. Using a case study, we demonstrated how WordBias can be used as an *auditing tool* by data scientists to probe for different kinds of social biases. Furthermore, the comments from the domain experts pointed at its possible utility for students and researchers. Given that WordBias does not require any programming expertise and can be easily accessed via a Web Browser, it can serve as an *educational tool* for students and non-experts to learn how AI (word embedding model) might be plagued with multiple kinds of social biases. For researchers, our tool can expedite the bias discovery process by acting as a quick alternate to writing code. Future work might involve students, researchers and data scientists to further refine and evaluate the usability and utility of our tool for different target audiences.

Group Words. They play a critical role in computation of bias scores [48]. In our case, we have used group words which have been proposed in existing literature (see Appendix) to kick off the bias exploration process. The user is advised to examine the default set of group words and update them via the visual interface as required [2]. If the user chooses to add a new bias type (axis), they should choose the words carefully to get an accurate picture. So far, there is no objective way to choose group words. However, our tool can assist in selecting the most relevant group words by facilitating comparison against a set of alternatives (as recommended in [2]). Lets say the user wants to add a new axis for 'political orientation' and they have multiple set of group words to choose from. In such a case, the user can add multiple axes corresponding to each set of group words. Thereafter, the user can explore and compare bias scores for different words across these axes. Group words corresponding to the axis which best aligns with the user's domain knowledge can be chosen.

Word Embedding. We have focused on static word embeddings trained on an English language corpus (word2vec). Similar social biases based on gender, etc. have been found in embeddings trained on other languages like French, Spanish, Hindi, German, Arabic, Dutch, etc. [27, 31, 32, 36, 50]. Furthermore, contextualized word embeddings like BERT [15], Elmo [34], etc. have also been found to contain social biases based on gender, etc. [41, 44]. Future work will involve extending support for contextualized word embeddings and embeddings trained on other languages.

8 CONCLUSION

In this work, we designed, implemented and evaluated a novel visual interactive tool to discover intersectional biases in word embeddings. We demonstrated how our tool helped uncover biases against multiple intersectional groups encoded in Word2Vec embedding. The source of such biases can be training data, word embedding model or they might be false positives due to limitations of the bias metric or sub optimal group words. Future research might investigate the exact cause of such biases and develop effective counter strategies.

ACKNOWLEDGMENTS

We thank all domain experts for their time and feedback. This work was funded through NSF award IIS 1941613.

REFERENCES

[1] Yongsu Ahn and Yu-Ru Lin. 2019. Fairsight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1086–1095.
[2] Maria Antoniak and David Mimno. 2021. Bad Seeds: Evaluating Lexical Methods for Bias Measurement. In *Proceedings of ACL*.
[3] Amir Bakarov. 2018. A Survey of Word Embeddings Evaluation Methods. *arXiv preprint arXiv:1801.09536* (2018).
[4] 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. In *Advances in Neural Information Processing Systems*. 4349–4357.
[5] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.
[6] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. *arXiv preprint arXiv:1904.05419* (2019).
[7] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[8] Shenghui Cheng and Klaus Mueller. 2015. The data context map: Fusing data and attributes into a unified display. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 121–130.
[9] Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality*. John Wiley & Sons.
[10] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.* (1989), 139.
[11] Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.* 43 (1990), 1241.
[12] Kimberlé W Crenshaw. 2017. *On intersectionality: Essential writings*. The New Press.
[13] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. 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*. 120–128.
[14] Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*. 879–887.
[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[16] Crystal Marie Fleming. 2018. *How to be less stupid about race: On racism, white supremacy, and the racial divide*. Beacon Press.
[17] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.
[18] Negin Ghavami and Letitia Anne Peplau. 2013. An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly* 37, 1 (2013), 113–127.
[19] Martin Graham and Jessie Kennedy. 2003. Using curves to enhance parallel coordinate visualisations. In *Proceedings on Seventh International Conference on Information Visualization, 2003. IV 2003.* IEEE, 10–16.
[20] Wei Guo and Aylin Caliskan. 2020. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. *arXiv preprint arXiv:2006.03955* (2020).
[21] Alfred Inselberg and Bernard Dimsdale. 1990. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proc. IEEE Visualization*. 361–378.
[22] Kai Chang. Accessed October 2020. Progressive Rendering. http://bl.ocks.org/syntagmatic/raw/3341641/ Progressive Rendering.
[23] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual analytics: Definition, process, and challenges. In *Information visualization*. Springer, 154–175.
[24] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional Bias in Hate Speech and Abusive Language Datasets. *arXiv preprint arXiv:2005.05921* (2020).
[25] Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84, 5 (2019), 905–949.
[26] Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics* 8 (2020), 486–503.
[27] Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional Analysis of Biases in Arabic Word Embeddings. *arXiv preprint arXiv:2011.01575* (2020).
[28] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047* (2019).
[29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[31] Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating Bias In Dutch Word Embeddings. *arXiv preprint arXiv:2011.00244* (2020).
[32] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 446–457.
[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
[34] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word

representations. *arXiv preprint arXiv:1802.05365* (2018).
[35] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* (2019), 1–19.
[36] Arun K Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing Gender biased Hindi Words with Word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*. 450–456.
[37] David Rozado. 2020. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PloS one* 15, 4 (2020), e0231189.
[38] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469* (2016).
[39] Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1679–1684.
[40] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 305–311.
[41] Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*. 13209–13220.
[42] Francisco Vargas and Ryan Cotterell. 2020. Exploring the linear subspace hypothesis in gender bias mitigation. *arXiv preprint arXiv:2009.09435* (2020).
[43] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation. *arXiv preprint arXiv:2005.00965* (2020).
[44] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.
[45] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.
[46] Wikipedia contributors. 2020. Percentile rank — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Percentile_rank&oldid=954713866 [Online; accessed 3-October-2020].
[47] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[48] Haiyang Zhang, Alison Sneyd, and Mark Stevenson. 2020. Robustness and Reliability of Gender Bias Assessment in WordEmbeddings: The Role of Base Pairs. *arXiv preprint arXiv:2010.02847* (2020).
[49] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).
[50] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. 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)*. 5279–5287.

A PREPROCESSING WORD EMBEDDING

Before loading the word embedding onto WordBias, we did some prepossessing similar to what is followed in the literature [4]. We only considered words with all lower case alphabets and whose length is upto 20 characters long. We then sorted the resulting words by their frequency in the training corpus and picked the most frequent 50,000 words. We made sure to include group words like names, etc. if they don't make it in the final list.

B FEATURE SCALING

WordBias allows the user to choose between raw bias scores and two feature scaling methods namely, Min-Max Normalization and Percentile Ranking. Raw bias scores provides the most accurate representation but it can be a bit difficult to interpret. The other two feature scaling options makes the bias scores more comparable across bias types. Figure 4 shows the distribution of mean bias scores for all 3 options. As we can see, the distribution of bias scores appear similar for (a) and (b) but their ranges on x-axis vary. This is because Min-Max normalization simply stretches the raw bias scores over the range [-1,1]. This figure also suggests that a large majority of words have small bias scores and only a few words on either ends have high bias scores. The distribution for Percentile ranking (Figure 4 (c)) is quite different and interesting. It has the same range on x-axis [-1,1] as Min-Max normalization but the distribution of words across bias scores is much more uniform. We can observe the the bar length is different for bias scores greater than and less than 0. This is because we applied percentile ranking in a piece-wise fashion depending on the sign of the bias scores. Figure 5 further elucidates the difference in distribution of bias scores for Min-Max normalization and Percentile ranking.

C GROUP WORDS

By default, WordBias shows 5 kinds of biases namely Gender, Religion, Age, Race and Economic. Following are the list of words used to compute bias scores for each of those categories. These words are derived from existing literature [14, 17, 25]. If any of these words aren't contained in the word embedding, they are ignored.

Male (Gender) [17]
he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews

Female (Gender) [17]
she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, sisters, aunt, aunts, niece, nieces

Young (Age) [14]
Taylor, Jamie, Daniel, Aubrey, Alison, Miranda, Jacob, Arthur, Aaron, Ethan

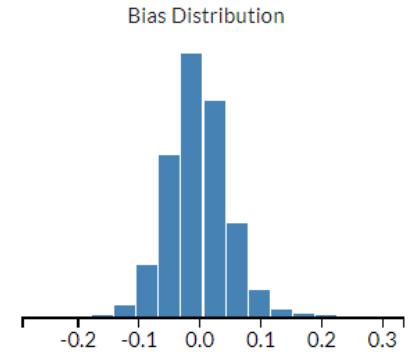
Old (Age) [14]
Ruth, William, Horace, Mary, Susie, Amy, John, Henry, Edward, Elizabeth

Islam (Religion) [17]
allah, ramadan, turban, emir, salaam, sunni, koran, imam, sultan, prophet, veil, ayatollah, shiite, mosque, islam, sheik, muslim, muhammad

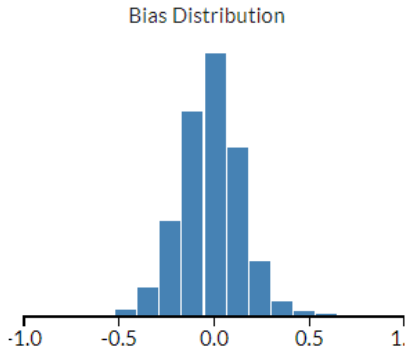
Christainity (Religion) [17]
baptism, messiah, catholicism, resurrection, christianity, salvation, protestant, gospel, trinity, jesus, christ, christian, cross, catholic, church

Black (Race) [25]
black, blacks, Black, Blacks, African, african, Afro

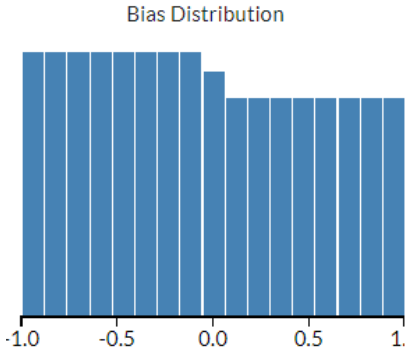
White (Race) [25]
white, whites, White, Whites, Caucasian, caucasian, European, european, Anglo



(a) Raw bias scores

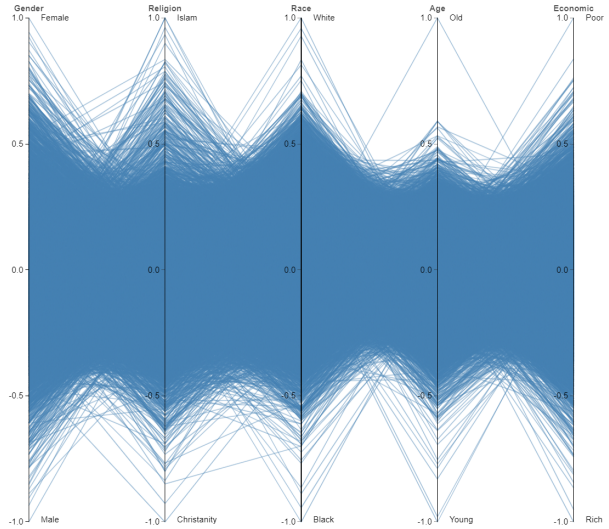


(b) Min-Max Normalization

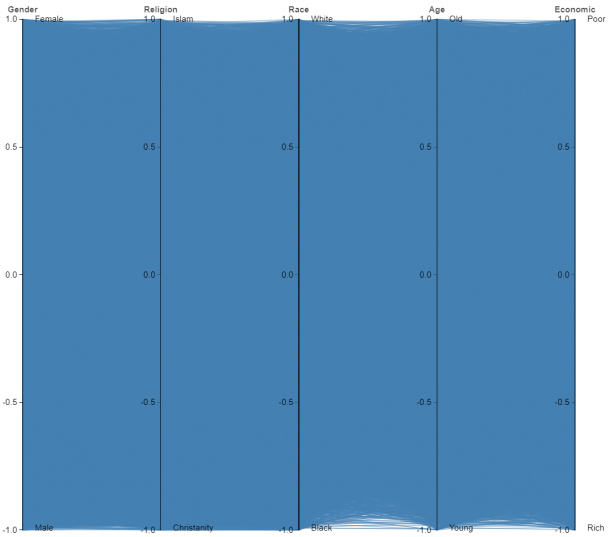


(c) Percentile Ranking

Figure 4: Distribution of bias scores across 50k words in the Word2Vec Embedding.



(a) Min-Max Normalization



(b) Percentile Ranking

Figure 5: Parallel coordinate plot for 50k words in the Word2Vec Embedding.

Rich (economic) [25]
rich, richer, richest, affluence, advantaged, wealthy, costly, exorbitant, expensive, exquisite, extravagant, flush, invaluable, lavish, luxuriant, luxurious, luxury, moneyed, opulent, plush, precious, priceless, privileged, prosperous, classy

Poor (economic) [25]
poor, poorer, poorest, poverty, destitude, needy, impoverished, economical, inexpensive, ruined, cheap, penurious, underprivileged, penniless, valueless, penury, indigence, bankrupt, beggarly, moneyless, insolvent

D NEUTRAL WORDS

To quickly audit a given embedding for different biases, WordBias provides a set of words which should ideally be neutral for most kinds of biases like gender, race, etc. Following is the list of such neutral words based on different categories which are derived from existing literature [7, 17].

Profession
teacher, author, mechanic, broker, baker, surveyor, laborer, surgeon,

gardener, painter, dentist, janitor, athlete, manager, conductor, carpenter, housekeeper, secretary, economist, geologist, clerk, doctor, judge, physician, lawyer, artist, instructor, dancer, photographer, inspector, musician, soldier, librarian, professor, psychologist, nurse, sailor, accountant, architect, chemist, administrator, physicist, scientist, farmer

Physical Appearance

alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong

Extremism

terror, terrorism, violence, attack, death, military, war, radical, injuries, bomb, target,conflict, dangerous, kill, murder, strike, dead, violence, fight, death, force, stronghold, wreckage, aggression,slaughter, execute, overthrow, casualties, massacre, retaliation, proliferation, militia, hostility, debris, acid,execution, militant, rocket, guerrilla, sacrifice, enemy, soldier, terrorist, missile, hostile, revolution, resistance, shoot

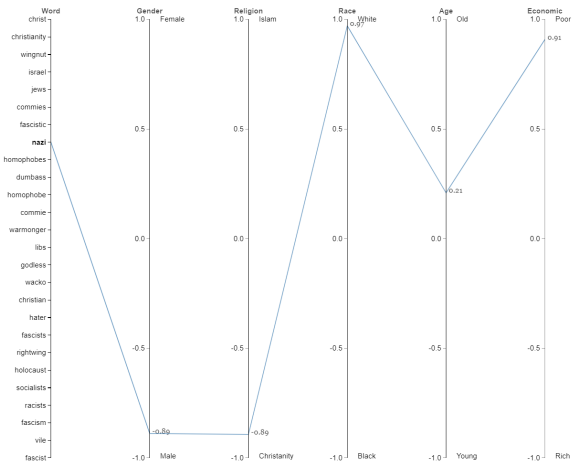
Personality Traits

adventurous, helpful, affable, humble, capable, imaginative, charming, impartial, confident, independent, conscientious, keen, cultured, meticulous, dependable, observant, discreet, optimistic, persistent, encouraging, precise, exuberant, reliable, fair, trusting, fearless, valiant, gregarious, arrogant, rude, sarcastic, cowardly, dishonest, sneaky, stingy, impulsive, sullen, lazy, surly, malicious, obnoxious, unfriendly, picky, unruly, pompous, vulgar

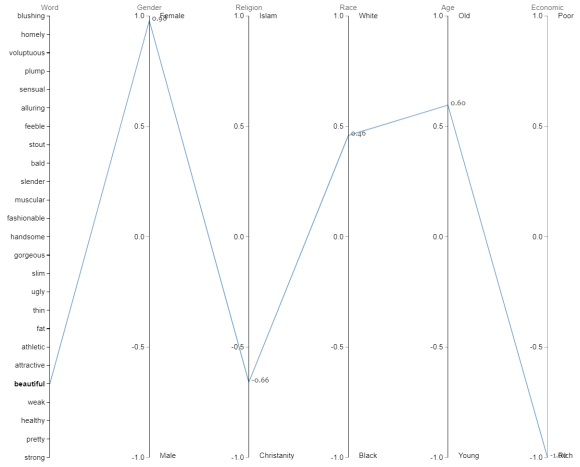
E FEW EXAMPLES

In the following, we list a few words along with their associated subgroups as per WordBias. Here, we have chosen percentile ranking and considered an association significant if its corresponding bias score is >= 0.5.

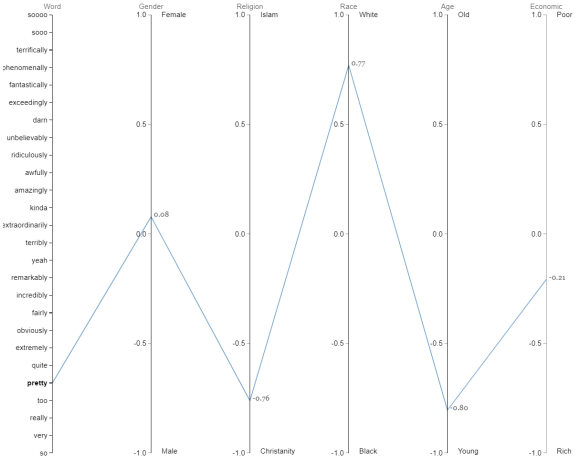
(i) nazi : Male - Christianity - White - Poor



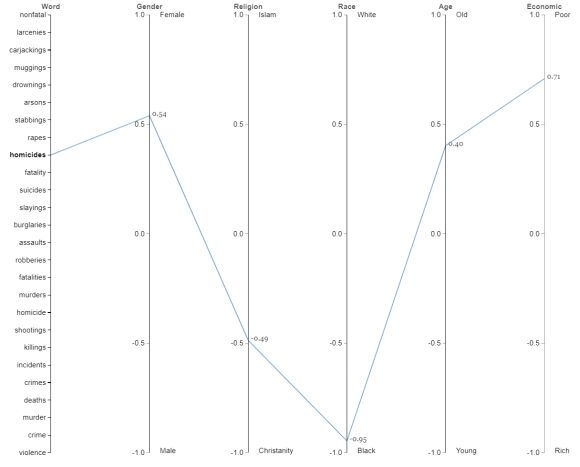
(ii) beautiful : Female - Christianity - Old - Rich



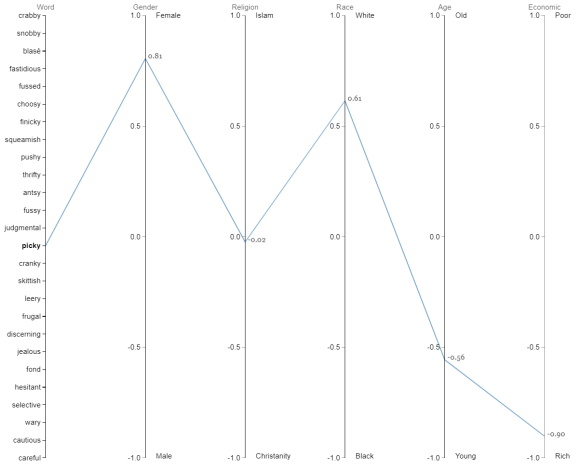
(iii) pretty : Christianity - White - Young



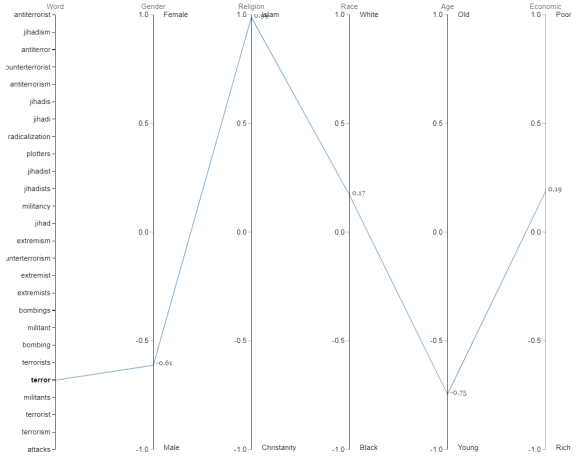
(iv) homicides : Female - Black - Poor



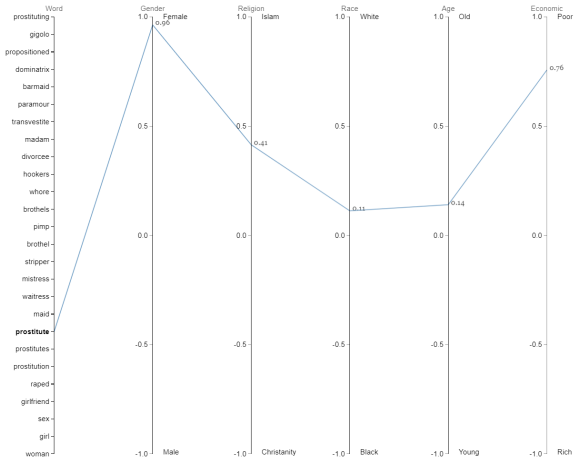
(v) picky : Female - White - Young - Rich



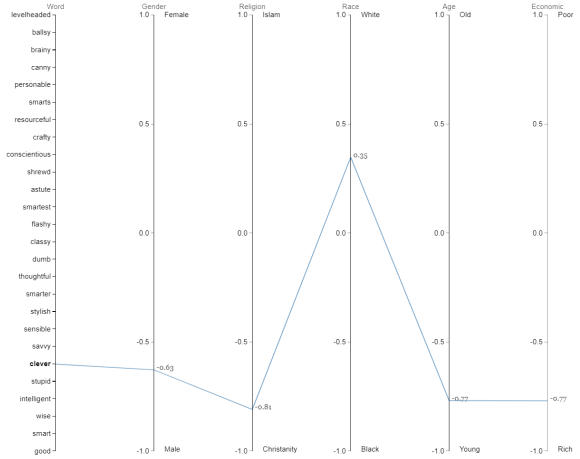
(vi) terror : Male - Islam - Young



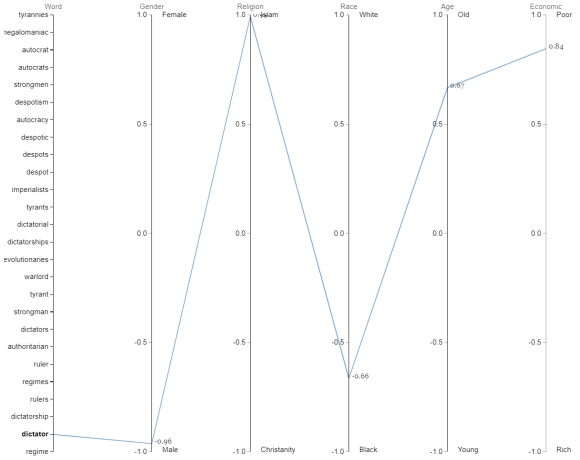
(vii) prostitute : Female - Poor



(viii) clever : Male - Christianity - Young - Rich



(ix) dictator : Male - Islam - Black - Old - Poor



(x) janitor : Male - Old - Poor