Exploring Differences in Word Associations across Online Communities

Jina Suh **Emil Azadian** **Galen Weld**
University of Washington
Paul G. Allen School of Computer Science and Engineering
{jinasuh,emilaz,gweld}@cs.washington.edu

Abstract

Word embeddings are known to capture associations that may reflect stereotypes present in the training corpus. While these stereotypical associations might be undesired in some contexts, they could yield an excellent tool for inspecting existing semantic associations within online communities. In this paper, we use word embeddings trained on texts from three online communities to explore how the same word is used and associated differently across the communities. We examine four different approaches to define a bias concept dimension, explore the embeddings using an interactive visualization tool, and introduce skewness as a metric to quantify the degree of bias.

1 Introduction

Word embedding is a natural language processing (NLP) technique that maps words (or phrases, subwords, or even characters) to vectors. Availability of toolkits and pre-trained embeddings(e.g., word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014)) have popularized the use of the embeddings in many downstream tasks such as text sentiment classification, part-of-speech tagging, and name-entity recognition. One of the main strengths of word embeddings is that it captures semantic relationships between words well. For example, given the vectors for 'king', 'man' and 'woman', the result of 'king' - 'man' + 'woman' $\approx$ 'queen' (Mikolov et al., 2013b).

Word embeddings are also known to capture associations that may reflect stereotypes present in the training corpus (Bolukbasi et al., 2016; Caliskan et al., 2017) to produce analogies such as 'computer programmer' - 'man' + 'woman' = 'homemaker.' Such undesirable associations can be shown to propagate to downstream tasks. Easily accessible pre-trained embeddings are used already in many existing tasks, such as sentiment analysis. In a recent analysis by Kiritchenko and Mohammad, 75% to 85% of the 219 sentiment analysis systems consistently predicted sentiments on sentences related to one gender differently from those of the other gender (Kiritchenko and Mohammad, 2018), demonstrating the gender imbalance in many existing systems today. Some have proposed methods to 'fix' the word embeddings (Bolukbasi et al., 2016) while others have argued against or criticized fixing it (Schnabel et al., 2015; Gonen and Goldberg, 2019). It seems more likely that taking off-the-shelf word embeddings or irresponsibly relying on embedding techniques without inspection or mitigation could lead to human harm.

Figure 1: Word co-occurrence clouds for the word 'oven' produced from comments on Gab, on the left, and Reddit, on the right. (Baumgartner, 2018)

While these stereotypical associations might be undesired in some contexts, they could yield an excellent tool for inspecting existing semantic associations within online communities, since words often have different connotations in different contexts, depending on the community using them. For example, Gab is an online social network that is known to attract alt-right users and conspiracy theorists (Zannettou et al., 2018). An analysis of comments on two social media platforms, Gab and Reddit, found dramatically different words to be used in conjunction with the word 'oven,' displayed in Figure 1. The popular website reddit consists of many subreddits, online communities focused around a specific theme or topic. Since

these groups attract distinct user bases, opinions, and topics, the language and use of words within these groups might significantly differ from, say, that of Wikipedia. In this project, we aim to use word embeddings as a tool to capture semantics of words, and explore these differences in semantics across online communities.

2 Related Work

2.1 Unwanted Associations in Word Embeddings

Some work has been done on methods for detecting, and potentially reducing, learned bias in word embeddings. Caliskan et al. show that bias in embeddings mirrors the human biases present in the corpora the embeddings are trained upon (Caliskan et al., 2017). Bolukbasi et al. investigate female/male stereotypes in the word2vec embedding trained on the Google News corpus, and propose a debiasing framework (Bolukbasi et al., 2016). Swinger et al. demonstrate a highly-unsupervised bias detection system on a number of publicly available embeddings (Swinger et al., 2018).

2.2 Semantic Differences

Words have different meanings in different contexts or times, and word embeddings could be used to capture the dominant meaning of the word in the training corpus. Hamilton et al. (Hamilton et al., 2016) have looked at historical semantic evolution; by aligning word embeddings across different time periods, the authors were able to visualize the change paths in context of other words. Rather than looking at how semantics evolve over time, our project will focus on how semantics of everyday words differ across online communities.

2.3 Visualization and Interaction Support

Interactive visualizations and interfaces have been useful in analyzing word embedding models because the nature of the problems that word embeddings address are human-centric and benefit largely from domain expertise or human supervision in the analysis process (Heimerl and Gleicher, 2018). There have been advances in algorithmic techniques for projecting high-dimensional embeddings into 2D space (e.g., tSNE (Maaten and Hinton, 2008)) with associated visualization support (Smilkov et al., 2016)[1]. Some interactive tools are designed to facilitate exploring and inspecting word embeddings alone (e.g., (Liu et al., 2018; Rong, 2014; Rong and Adar, 2016)) while others support a related user task (e.g., interactive lexicon building (Park et al., 2018)). Heimerl and Gleicher (Heimerl and Gleicher, 2018) summarize linguistic tasks that word embeddings employ and the role of visualization in support of those tasks. Our project focuses on facilitating human exploration where the user is engaged in defining and comparing concepts (e.g., gender, race) and understanding the differences in the projection of the same word along a concept axis across different datasets.

3 Data and Word Embedding Models

In order to train our word embedding models, we prepared our dataset from three online sources: (1) One Billion Word Language Model Benchmark[2] dataset provides a large, high-quality 'reference' corpus to compare other communities' language against and consists of news articles from various online news sources. (2) Reddit[3] is a social media and discussion site with over 500 million monthly users, making it the 3rd most popular website in the US. Reddit's users subscribe to and post on various topics or themes. Our Reddit dataset consists of randomly sampled posts at 1%. (3) Gab[4] is a social media website advertised as a platform for free-speech, making it a safe-haven for white supremacists, neo-Nazis, and alt-rights. Our Gab dataset represents online discussion language from such predominantly alt-right community.

For Reddit and Gab data, where each post could span multiple sentences or paragraphs, we used `Punkt` sentence tokenizer from Natural Language Toolkit (`nltk`). We then tokenized all sentences using `TweetTokenizer`, designed to handle the kinds of tokens often found in social media posts. We filtered all sentences with less than 3 tokens that are often short responses or reactions and do not carry much information about the associations between words used in a sentence. For each dataset, we built skip-gram word2vec models of length 300 using `gensim` library. We allowed unigrams and bigrams with minimum occurrence of 5 in our vocabulary (see Table 1). The resulting

[1] http://projector.tensorflow.org/

[2] http://www.statmt.org/lm-benchmark/

[3] https://files.pushshift.io/reddit/comments/; from May 2018 to October 2018

[4] https://files.pushshift.io/gab/; from August 2016 to October 2018

Dataset	Tokens/Sentence	Vocabulary
One Billion	25.4	751,333
Reddit	16.6	363,986
Gab	14.0	1,255,273

Table 1: The number of tokens per sentence and the size of the vocabulary for three trained word embeddings.

three word embeddings shared $110,073$ words in the common vocabulary.

4 Bias Concept Dimensions

Our goal is to understand the differences in the use of words and the contexts in which the words appear across different online communities. Rather than comparing individual words holistically, we want to define a *bias concept dimension*. With a bias concept dimension, we can characterize the association or relatedness of each word along that dimension. Since both the words and dimensions are represented by d-dimensional vectors in the word embedding space, we can compute such association through simple techniques such as *cosine similarity*. Cosine similarity of a word to a concept dimension can be computed by

$$cos(\overrightarrow{w}, \overrightarrow{D_c}) = \frac{\overrightarrow{w} \cdot \overrightarrow{D_c}}{|\overrightarrow{w}||\overrightarrow{D_c}|}$$

where $\overrightarrow{D_c}$ is a vector representing the bias concept dimension and $\overrightarrow{w}$ is the word vector.

A concept dimension could take many forms (Park et al., 2018): a *bipolar* dimension has two concepts that may oppose or distinguish from each other (e.g., night and day, man and woman), and a *unipolar* dimension has one concept and measures the degree of relatedness or presence of that concept (e.g., cooking-related). In this work, we focus on defining a bipolar concept dimension to explore the dichotomy between stereotypical bias pairs through the lens of how other words are associated along that concept dimension.

A bias concept dimension in the word embedding space can be defined by a pair of word vectors (e.g., man and woman). However, because words have multiple meanings and usages, a single word (e.g., man) may not precisely represent the whole concept. Bolukbasi et al. (Bolukbasi et al., 2016) suggested a more robust way to define a concept dimension or subspace by aggregating across multiple paired comparison using *definitional word pairs*. Definitional word pairs are pairs of words that describe the two contrasting concepts. For example, "he-she", "male-female", "man-woman" are pairs of words that describe the gender dichotomy. Bolukbasi et al. define a gender subspace by computing principal components (PCs) of ten gender pair difference vectors (e.g., $\overrightarrow{man} - \overrightarrow{woman}$). In our work, we leverage their idea of defining a bias concept dimension using principal component analysis (PCA) of a set of definitional word pairs and experiment with three additional variants to their approach.

4.1 Best Bias Concept Dimension

Given a set of definitional pairs $\{(w_i, v_i)\}_{i=1}^{n}$, we define the accuracy of each definitional pair as follows. For each pair (w_j, v_j), we compute the pair dimension by taking the difference of the two word vectors, $\overrightarrow{w_j} - \overrightarrow{v_j}$. We project every word in the definitional pairs onto this axis by computing the cosine similarity of the word with the axis, $cos(\overrightarrow{w_i}, \overrightarrow{w_j} - \overrightarrow{v_j})$. We then count the number of words that appear on the *correct side* of the axis. For example, for $\overrightarrow{man} - \overrightarrow{woman}$ axis, the projection of the word "he" should be closer to "man", and therefore, we should see $cos(\overrightarrow{he}, \overrightarrow{man} - \overrightarrow{woman}) > 0$. By repeating this check for every word in the definitional pairs, we can compute per-pair accuracy as

$$\text{Accuracy} = \frac{\text{Number of words correctly placed}}{\text{Total number of words}}$$

The quality of the bias concept dimension depends on the qualities of the definitional pairs as well as the characteristics of each word embedding. We have four approaches for selecting the subset of the definitional pairs that best fit the other pairs: (1) *All* uses all given definitional pairs, and this method is introduced and used in Bolukbasi et al. (Bolukbasi et al., 2016). (2) *Best for each* chooses the top N definitional pairs whose accuracy is the highest for each word embedding. (3) *Best for all* chooses the top N definitional pairs whose average accuracy across the word embeddings is the highest. (4) *Best by CV* splits the pairs in a k-fold manner, takes the first principal direction computed on the current fold and calculates the accuracy over all test words on that principal direction. The principal direction with the highest accuracy is then set as the bias concept dimension.

	Labeled							
	All		**Best each**		**Best all**		**CV**	
	M	F	M	F	M	F	M	F
M	111	21	118	14	115	17	113	19
F	3	129	3	129	4	128	7	125
Acc.	0.909		0.936		0.920		0.902	

Table 2: Confusion matrices and accuracies for four different approaches of selecting 5 best definitional pairs out of 44 across 3 word embeddings. The rows present predicted concepts (i.e., male or female) and the columns present labeled concepts.

For (1)-(3), using the chosen N definitional pairs, we then perform a PCA and choose the first PC dimension to be the bias concept dimension.

4.2 Evaluation of Concept Dimensions

To evaluate our approaches for computing the best bias concept dimension, we took 52 definitional pairs describing "man-woman" concepts from Bolukbasi et al. (Bolukbasi et al., 2016) (see Table 5 in the Appendix). We excluded capitalized words and out-of-vocabulary words, and we also modified "Catholic_priest" to "priest" to be balanced with its pair "nun". In the end, we were left with 44 pairs. We used the four approaches described above to select 5 best definitional pairs out of the given 44 pairs. Table 2 shows the confusion matrices and accuracies of the four approaches. *Best for each* method performed the best in accurately placing the words.

4.3 Word association scores

Once the bias concept dimension is defined, we compute a *word association score* as cosine similarity between the word and the newfound dimension. While this score cannot be directly translated into the word's actual tendency towards one concept over another, it describes a relative tendency, especially when one score is compared with that of another word. For example, given a "male-female" concept dimension within a single word embedding, a word with a score 0.3 is more closely associated with "male" than a word with a score of 0.2 or -0.1. A word with a score of -0.3 is more closely associated with "female" than a word with a score of -0.1. A complicating factor for making comparisons of these scores across different embeddings is the fact that the lengths and positions of word vectors carry not absolute meaning and obtain their informativeness solely from their relative placement within an embedding. In an attempt to make the word association scores more comparable, we normalize the scores by the maximum absolute word association score within an embedding. This way, we are looking at the relative differences between words rather than their absolute differences.

5 Analysis

Based on our earlier analysis on defining bias concept dimensions, we analyzed the three word embeddings using both qualitative and quantitative approaches.

5.1 Qualitative: Interactive Visualization

To perform qualitative evaluation, we developed an interactive visualization that allows a user to explore the differences across the three word embeddings. The interactive visualization supports three main tasks which can be seen in Figure 2.

5.1.1 Define bias concept dimensions

Users can enter their own definitional word pairs to define their own bias concept dimensions (Figure 2A). They can edit an existing dimension at any time as they discover relevant terms to add or to remove unhelpful words. They can also create a new dimension as they find new concepts to explore. Users can switch between concept dimensions to focus on different concepts.

5.1.2 Inspect interesting words

Inspection of the word associations with the selected concept dimension is visualized using a parallel coordinate (Figure 2C). Each word embedding displays a concept axis and each word is plotted on the axis based on its word association score described earlier. Rather than plotting the word associations for the entire common vocabulary, we provide four different ways of filtering (Figure 2B) so that the user can focus their attention on potentially interesting or surprising words: (1) *Most variance* chooses top N words with maximum variance in the word association scores, (2) *Most pairwise difference* chooses top N words with maximum pair-wise differences in the word association scores, (3) *Most opposite* chooses top N words from each embedding that are most similar to each extreme end of the dimension, and (4) *Dimension* chooses words in the definitional pairs.

Most variance and *most pairwise difference* filters can suggest words that characterize the most

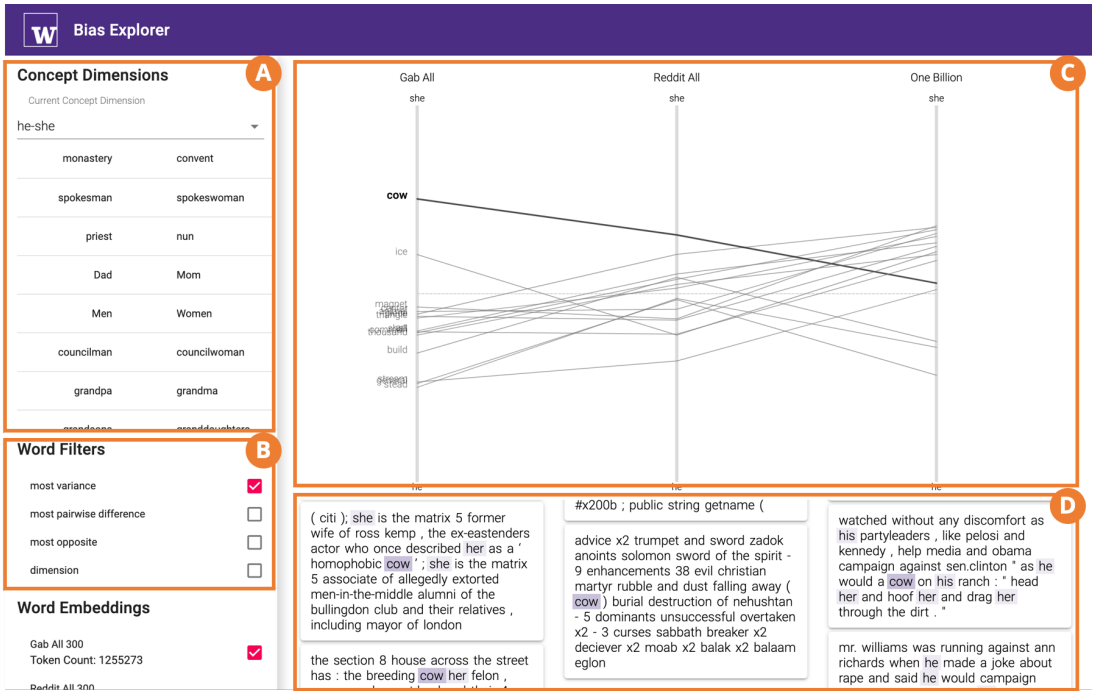


Figure 2: Interactive visualization used to explore the word embeddings and various word associations on concept dimensions. A user can define their own concept dimensions (A), select filters to focus on specific words (B), see the word associations along the concept dimension across the embeddings (C), and examine contexts in which the selected words co-occur with the words that define the concepts (D).

striking differences among the word embeddings. *Most opposite* and *dimension* filters allow the users to evaluate the quality of the dimensions across word embeddings.

5.1.3 Context Exploration

When a user selects a word to inspect, we display a set of example sentences from each corpus (Figure 2D). These example sentences help the user to better understand the context in which a word is used, and how these contexts vary from corpus to corpus - how their usage varies from online community to community. To achieve this goal, example sentences selected must be somewhat representative of the word embedding's geometry surrounding the word selected by the user. To achieve this goal, we preprocess each sentence in the corpora, mapping it to the words it contains, and then in real-time select sentences that contain the greatest number of occurrences of target words and definitional pair words. By showing the contexts, the users can validate their understanding about the word associations for the particular word embedding or compare the usage of words across different datasets from which the embeddings were trained.

The example context generation feature, while the most qualitative of all modes of analysis used, immediately permits the user to read real uses of a given word, and see how they vary across corpora.

Consider uses of the word 'cow.' In the One Billion corpus, taken from news articles, instances of cow that co-occur with gendered words such as 'he,' 'she,' etc. tend to be related to agriculture:

`Mr. Williams said he would campaign against Ms. Richards as he would a cow on his ranch.`

In Gab, however, the examples are much worse.

`The house across the street has: the breeding cow, her felon gang-member not-husband, [and] their 4 children.`

These patterns hold for more than just words that can be used in an offensive manner. In the One Billion dataset, many instances of the word 'trump' refer to the verb:

`Health-care and environmental issues trump such traditional issues as equal pay, abortion rights and contraception coverage.`

Whereas in the Gab dataset, every usage of the word 'trump' refers to the person:

`What trump did for me was very`

`simple: he began the process of dismantling globalism.`

While these examples do not provide the same high-level picture of word differences that our visualization does, but they are still enormously valuable in providing the user with an understanding of how a word is used, in context.

5.2 Quantitative: Skewness

In order to quantify the degree of bias expressed in the word embeddings, we use a set of 40 professions that were used in (Zhao et al., 2018) and are originally pulled from the Labor Force Statistics [5]. Only one profession was not in the commonly vocabulary across the three datasets and was therefore not used, leaving us with 20 predominantly female and 19 predominantly male jobs. The full list can be found in Table 6 in the Appendix. We explore two naive ways of ascribing a gender skewness score to a dataset. First, we quantify skewness relative to real-life imbalances as reported in the Labor Force Statistics as follows:

$$\text{Relative Skewness} = \frac{1}{N}\sum_{i=1}^{N}\frac{|cos(\overrightarrow{w_i}, \overrightarrow{D_c})|}{r(w_i)}$$

where $cos(\overrightarrow{w_i}, \overrightarrow{D_c})$ is the word association score as introduced in Chapter 4 and $r(w_i)$ is the deviation from a balanced 50% gender employment in the Labor Force Statistics for a given profession w_i. We scale r to be in $[0, 1]$ to be comparable with the cosine similarity. The results can be found in 3. Note that relative skewness does not capture whether our word association score is imbalanced towards the opposite or same gender. Interestingly, we find that the datasets One Billion and Reddit capture current existing gender biases much more accurately than Gab, which is relatively less biased on the given professions.

As second metric, we report the confusion matrices as indicator on how well the direction of skewness in our word embeddings represent the actual imbalances in the job market. Looking at table 4 reveals that the One Billion reference corpus in fact best captures existing employment imbalances ($ACC = 79.5\%$), whereas Gab ($ACC = 74.4$) and Reddit ($ACC = 69.2$) are progressively worse at capturing these imbalances. A more in-depth discussion of this result will follow in the next section.

[5] https://www.bls.gov/cps/cpsaat11.htm

Relative Skewness

Dataset	Female	Male	Avg
One Billion	0.54	0.98	0.90
Reddit	0.81	0.67	0.73
Gab	0.58	0.49	0.53

Table 3: The skewness of the three datasets relative to actual gender imbalances in the selected professions. The columns Female/Male refer to jobs in which predominantly women/men are employed according to the Labor Force Statistics.

	One Billion		Reddit		Gab	
	M	F	M	F	M	F
Proj. M	18	7	15	8	15	6
Proj. F	1	13	4	12	4	14

Table 4: Confusion matrix for all three datasets. F/M refer to jobs with predominantly female/male employees (20F/19M), Proj. M/F refers to the word association score of those professions being closer to the male or female gender concept.

6 Discussion and Future Work

The primary contribution of this work is the visualization tool, which enables users to explore how a word's bias varies across different corpora from different online communities. By allowing for the user to select from one of a number of different concept dimensions, this tool goes beyond much existing analysis of bias in word embeddings, which mostly is limited to analysis of a single fixed concept dimension, frequently the he-she gender dimension.

The multi-faceted interactivity of the tool, which allows users to not only select a concept dimension, as described above, but also to select specific words to analyze, makes this tool a powerful one for building an intuition for different uses of words online. The examples presented for different words facilitate an understanding of what these different uses look like in their real-world contexts.

However, our system is not without limitations and flaws, and there is significant room for future improvement. A significant challenge for us stems from the observation that frequently, words' different uses originate in a given word's multiple meanings. For example, the word 'ice' is used quite differently in Gab versus in the One Billion corpus. As the examples make clear, uses of

'ice' on Gab tend to refer to ICE, the US government agency tasked with enforcing immigration law, whereas in the One Billion corpus, 'ice' more commonly refers to frozen water. It is interesting in and of itself to learn that words with multiple meanings, such as 'ice, ' are used more frequently with one meaning in one community than with a different meaning in a different community, but for a deeper understanding of language differences, we would like to be able to control for multiple word meanings.

This limitation is currently primarily a function of our word embedding training process, which produces a single embedding for each word, effectively an average of all of that word's different meanings. More modern word embedding models enable the creation of one embedding per word-meaning, as opposed to one embedding per word. This increased granularity would enable us to compare the differences between uses of (for example) 'ice' (as in frozen water) between Gab and the One Billion corpus.

Another challenge with our system, as it is currently implemented, is the prevalence of uncommon or foreign language words under various word filters. These word filters exist to suggest to the user which words are used differently in different corpora, but our methods for automatically selecting these words based on the word-embedding geometries often results in the appearance of words which are not meaningful to most English speakers. We would like to identify a dataset of commonly used English words, and use this as a filter, in order to increase the impact of the words presented to the user in the visualization.

Furthermore, we would like to increase the number of concept dimensions available to the user, and permit the user to define their own concept dimension by providing a set of antonym pairs. We would like to increase the number of available word embeddings available for the user to explore, for example, but offerering embeddings for a number of different subreddits. All of these goals increase the scope of the project, but make the system more valuable to the user.

The quantitative analysis of the online communities in question were able to reveal quite unexpected results. Using a word embedding of an alt-right community like GAB may be leading to the early conclusion of the community being more biased than neutral baselines like the One Billion dataset. However, we were able to show that Gab is in fact the least skewed dataset among the three when quantified on gender-imbalanced professions. We suspect that this effect might be due to larger datasets that are trained on a more balanced corpus capturing real life gender imbalances in professions (a) with more accuracy and (b) to a higher degree in terms of skewness. As open question remains the context of words and professions used. It might further be interesting the current results to even more niche online communities.

7 Conclusion

We've presented Bias Explorer, an interactive visualization system for exploring differences in language usage across different online communities. While most existing work on bias in word embeddings focuses on an analysis of a single, fixed concept, such as gender bias, our system allows the user to explore bias on a number of different concept dimensions. We have provided a naive way of comparing the degree of skewness across those communities and were able to show that more radical communities like GAB are not necessarily more skewed than more "neutral" ones.

Our initial prototype fulfills our major design goals and effectively allows a user to analyze and learn more about language bias. However, we're very excited about continuing to develop and further refine the system, improving the quality of our word embeddings, training on more corpora, and allowing the user to define their own concept dimensions.

Acknowledgments

We are deeply appreciative all the encouragement and advice we recieved from Professor Yejin Choi, and the TAs for CSE 517.

References

Jason Baumgartner. 2018. Two social media platforms' use of oven.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically

from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Florian Heimerl and Michael Gleicher. 2018. Interactive analysis of word vector embeddings. In *Computer Graphics Forum*, volume 37, pages 253–265. Wiley Online Library.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Shusen Liu, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. 2018. Visual exploration of semantic relationships in neural word embeddings. *IEEE transactions on visualization and computer graphics*, 24(1):553–562.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*.

Deokgun Park, Seungyeon Kim, Jurim Lee, Jaegul Choo, Nicholas Diakopoulos, and Niklas Elmqvist. 2018. Conceptvector: text visual analytics via interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics*, 24(1):361–370.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Xin Rong. 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Xin Rong and Eytan Adar. 2016. Visual tools for debugging neural language models. In *Proceedings of ICML Workshop on Visualization for Deep Learning*.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. *CoRR*, abs/1611.05469.

Nathaniel Swinger, Maria De-Arteaga, IV Heffernan, Neil Thomas, Mark DM Leiserson, and Adam Tauman Kalai. 2018. What are the biases in my word embedding? *arXiv preprint arXiv:1812.08769*.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, Haewoon Kwak, and Jeremy Blackburn. 2018. What is gab? a bastion of free speech or an alt-right echo chamber? *arXiv preprint arXiv:1802.05287*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876.

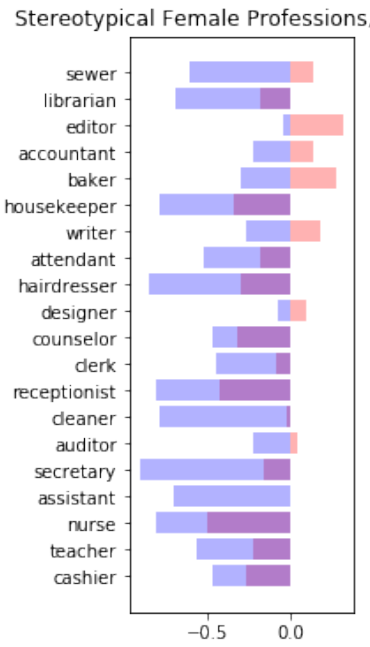
8 Appendix

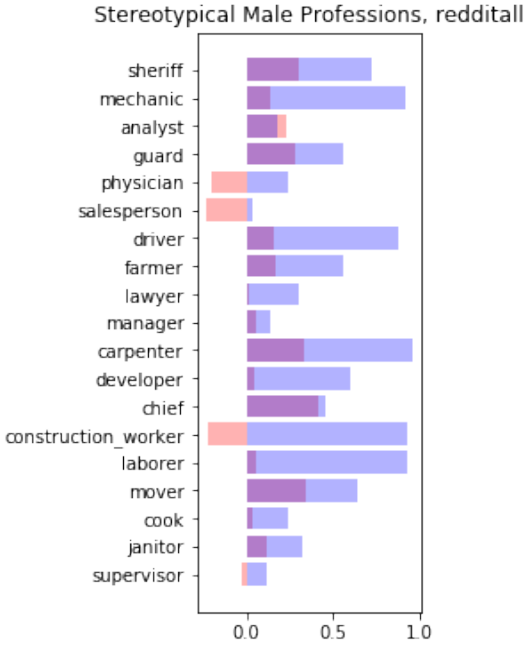
Male	Female	Male	Female
monastery	convent	*ex_girlfriend*	*ex_boyfriend*
spokesman	spokeswoman	father	mother
priest	nun	fatherhood	motherhood
Dad	*Mom*	fathers	mothers
Men	*Women*	fella	granny
councilman	*councilwoman*	fraternity	sorority
grandpa	grandma	*gelding*	*mare*
grandsons	granddaughters	gentleman	lady
prostate_cancer	ovarian_cancer	gentlemen	ladies
testosterone	estrogen	grandfather	grandmother
uncle	aunt	grandson	granddaughter
wives	husbands	he	she
Father	*Mother*	himself	herself
Grandpa	*Grandma*	his	her
He	*She*	king	queen
boy	girl	kings	queens
boys	girls	male	female
brother	sister	males	females
brothers	sisters	man	woman
businessman	businesswoman	men	women
chairman	chairwoman	nephew	niece
colt	filly	prince	princess
congressman	congresswoman	schoolboy	schoolgirl
dad	mom	son	daughter
dads	moms	sons	daughters
dudes	gals	twin_brother	twin_sister

Table 5: Definitional pairs used in the evaluation of concept dimensions (Bolukbasi et al., 2016). Italicized words were not part of the common vocabulary.

Male biased		**Female biased**	
supervisor	44	cashier	73
janitor	34	teacher	78
cook	38	nurse	90
mover	18	assistant	85
laborer	3.5	secretary	95
construction_worker	3.5	auditor	61
chief	27	cleaner	89
developer	20	receptionist	90
carpenter	2.1	clerk	72
manager	43	counselor	73
lawyer	35	designer	54
farmer	22	hairdresser	92
driver	6	attendant	76
salesperson	48	writer	63
physician	38	housekeeper	89
guard	22	baker	65
analyst	41	accountant	61
mechanic	4	editor	52
sheriff	14	librarian	84
(CEO)	39	sewer	80

Table 6: List of professions used for analyzing bias (Zhao et al., 2018). The numbers represent the percentages of women in those professions according to the Labor Force Statistics. CEO was not in the common vocabulary and therefore not used.


Stereotypical Female Professions, redditall
sewer
librarian
editor
accountant
baker
housekeeper
writer
attendant
hairdresser
designer
counselor
clerk
receptionist
cleaner
auditor
secretary
assistant
nurse
teacher
cashier
−0.5
0.0


Stereotypical Male Professions, redditall
sheriff
mechanic
analyst
guard
physician
salesperson
driver
farmer
lawyer
manager
carpenter
developer
chief
construction_worker
laborer
mover
cook
janitor
supervisor
0.0
0.5
1.0

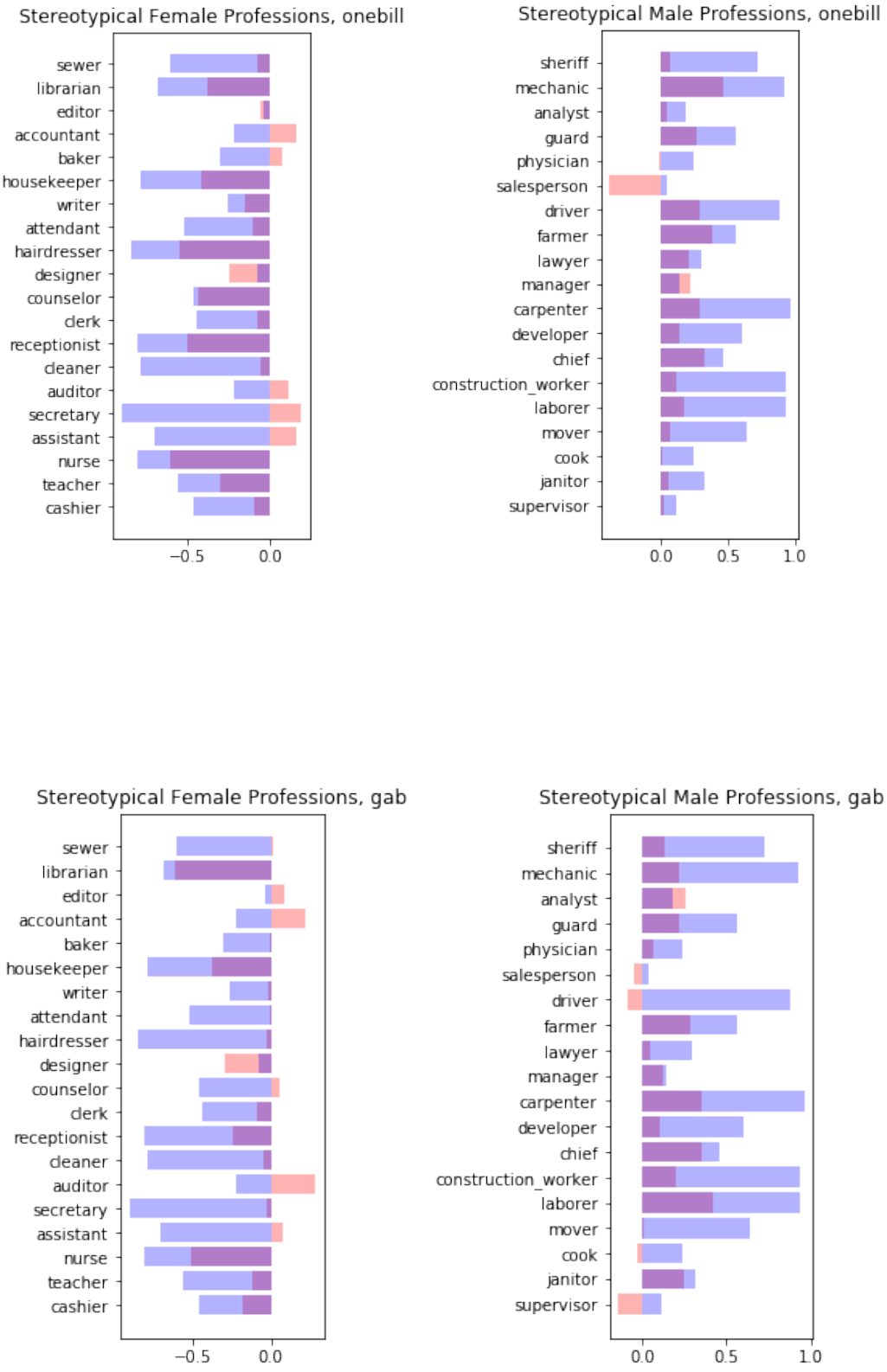


Figure 3: Differences in actual skewdness (blue) and word association score (red) for all three datasets.