# On the interpretability and significance of bias metrics in texts: a PMI-based approach

**Anonymous ACL submission**

## Abstract

In recent years, the use of word embeddings has become popular to measure the presence of biases in texts. Despite the fact that these measures have been proven to be effective in detecting a wide variety of biases, metrics based on word embeddings lack transparency, explainability and interpretability. In this study, we propose a PMI-based metric to quantify biases in texts. We prove that this metric can be approximated by an odds ratio, which allows estimating the confidence interval and statistical significance of textual bias. This PMI-based measure can be expressed as a function of conditional probabilities, providing a simple interpretation in terms of word co-occurrences. Our approach produces a performance comparable to GloVe-based and skip-gram-based metrics in experiments of gender-occupation and gender-name associations. We discuss the advantages and disadvantages of using methods based on first-order vs second-order co-occurrences, from the point of view of the interpretability of the metric and the sparseness of the data.

## 1 Introduction

Techniques for estimating the semantic closeness between words are a core component in a wide variety of NLP applications. One particularly prolific idea behind these methods is that the meaning of words is at least partially determined by the context in which they appear (Harris, 1954), of which word embeddings are their most popular instantiation (Mikolov et al., 2013; Pennington et al., 2014).

The flexibility and scalability of word embeddings have made them ideal in the study of textual biases (Bolukbasi et al., 2016; Hamilton et al., 2016; Kulkarni et al., 2015; DeFranza et al., 2020; Charlesworth et al., 2021). In particular, word embedding-based approaches have been used to detect and quantify the presence of gender, ethnic, racial and other stereotypes in texts (Lenton et al., 2009; Caliskan et al., 2017; Garg et al., 2018).

### 1.1 Bias quantification in texts

Consider two sets of context words *A* and *B*, and a set of target words *C*. Generally speaking, bias measures seek to quantify how much more the words of *C* are associated with the words of *A* than with those of *B* (or vice versa). Consider for instance the case of gender biases in occupations: the fact that certain jobs (C) are more likely to be associated with one particular gender (A) than other(s) (B). Here context words are often taken to be gendered pronouns or nouns, e.g., A = *{she, her, woman,..}* and *B* = *{he, him, man,...}* (Bolukbasi et al., 2016; Garg et al., 2018; Caliskan et al., 2017; Gálvez et al., 2019; Lenton et al., 2009; Lewis and Lupyan, 2020).

A popular choice for C is to consider one word at a time, i.e. estimating how a specific job (such as nurse, doctor or engineer) is associated with the two previously defined sets. Once these sets have been defined, most metrics can be parametrized as a difference between the similarities between A and C, on the one hand, and B and C, on the other (Lenton et al., 2009; Garg et al., 2018; Lewis and Lupyan, 2020); in most cases subtracting the similarities A vs C and B vs C:

$$Bias = sim(A, C) - sim(B, C), \qquad (1)$$

One particularly influential case that belongs to this class of metrics is that of Caliskan et al. (2017), who use

$$Bias = \frac{\underset{a \in A}{\text{mean}}(cos(v_a, v_c)) - \underset{b \in B}{\text{mean}}(cos(v_b, v_c))}{\underset{x \in A \cup B}{std\_dev}\ cos(v_x, v_c)}$$

where

$$cos(v_x, v_y) = \frac{v_x . v_y}{|v_x||v_y|},$$

and $v_i$ stands for the word embedding of word $i$.

In the present paper we propose a metric to measure biases that follows equation 1 based on

Pointwise Mutual Information (PMI) (Church and Hanks, 1990; Jurafsky and Martin, 2009) as our measure for similarity between words, so that

$$Bias_{PMI} = PMI(w_a, w_c) - PMI(w_b, w_c) \quad (2)$$

PMI is a non-parametric measure of statistical association derived from information theory (Fano, 1961). Given two words $w_x$ and $w_y$, PMI is defined as

$$PMI(w_x, w_y) = log\left(\frac{p(w_x, w_y)}{p(w_x)p(w_y)}\right) \quad (3)$$

$p(w_x, w_y)$ is the probability of co-occurrence of words $w_x$ and $w_y$ in a given window of $k$ words. The PMI (equation 3) compares the probability of co-occurrence of the words $w_x$ and $w_y$ with that of the case in which $w_x$ and $w_y$ are independent. It can also be used to calculate associations between word lists $X$ and $Y$:

$$PMI(X, Y) = log\left(\frac{p(X, Y)}{p(X)p(Y)}\right) \quad (4)$$

In equation 4, $p(X, Y)$ is the probability of co-occurrence between any word in $X$ with any one in $Y$. Similarly, $p(X)$ and $p(Y)$ are the probability of occurrence of any word in $X$ and any word in $Y$, respectively.

The usage of PMI in the study of textual biases is not new (Gálvez et al., 2019). In this article we explain the statistical and interpretability benefits associated with PMI-based measures, which have been overlooked until now.

In particular, we make the following contributions:

- In section 2 we show how the bias measurement using PMI can be approximated by an odds ratio. This comes with some statistical perks, such as the possibility of performing computationally inexpensive null hypothesis statistical testing.

- In section 3 we demonstrate that methods based on GloVe, skip-gram with negative sampling (SGNS) and PMI produce comparable results in Caliskan et al. (2017)'s tasks, in which real-world gender distributions of occupations and first names are compared with the biases measured in texts.

- In section 5 we contend that the use of PMI in bias metrics is substantially more transparent and interpretable than the counterparts based on word embedding techniques.

## 2 Approximation of the PMI-based bias metric by log odds ratio

The PMI between a list of context words $X$ and a list of target words $C$ can be expressed as the ratio between the probability of words in $C$ co-occurring in the context of words in $X$, and the probability of words in $C$ appearing in any context:

$$\begin{aligned} PMI(X, C) &= log\left(\frac{p(X, C)}{p(X)p(C)}\right) \\ &= log\left(\frac{p(C|X)}{p(C)}\right) \end{aligned}$$

Therefore, the PMI-based bias can be rewritten as follows:

$$\begin{aligned} Bias_{PMI} &= PMI(A, C) - PMI(B, C) \\ &= log\left(\frac{p(C|A)}{p(C)}\right) - log\left(\frac{p(C|B)}{p(C)}\right) \\ &= log\left(\frac{p(C|A)}{p(C|B)}\right) \end{aligned}$$

To compute the PMI-based bias, we can estimate probabilities via maximum likelihood:

$$Bias_{PMI} = log\left(\frac{\frac{f_{A,C}}{f_{A,C}+f_{A,nC}}}{\frac{f_{B,C}}{f_{B,C}+f_{B,nC}}}\right) \quad (5)$$

| | $C$ | *not* $C$ | total |
|---|---|---|---|
| $A$ | $f_{A,C}$ | $f_{A,nC}$ | $f_{A,C}+f_{A,nC}$ |
| $B$ | $f_{B,C}$ | $f_{B,nC}$ | $f_{B,C}+f_{B,nC}$ |

Table 1: Contingency Table of words co-occurrences

In equation 5, $f_{A,C}$ and $f_{B,C}$ represent the number of times words in $C$ appear in the context of words in $A$ and $B$, respectively, and $f_{A,nC}$ and $f_{B,nC}$ represent how many times words not in $C$ appear in the context of $A$ and $B$, respectively. See contingency table in Table 1 for reference.

In the case in which the following condition is fulfilled,

$$f_{B,nC} \gg f_{B,C},\ f_{A,nC} \gg f_{A,C} \quad (6)$$

equation 5 can be approximated by

$$\begin{aligned} Bias_{PMI} &= log\left(\frac{\frac{f_{A,C}}{f_{A,C}+f_{A,nC}}}{\frac{f_{B,C}}{f_{B,C}+f_{B,nC}}}\right) \\ &\approx log\left(\frac{\frac{f_{A,C}}{f_{A,nC}}}{\frac{f_{B,C}}{f_{B,nC}}}\right) \\ &\approx log(OR) \end{aligned}$$

where $OR$ is the odds ratio associated with Table 1. For large sample sizes, the distribution of $log(OR)$ converges to normality (Agresti, 2003). Therefore, this quantity has a 95% confidence interval given by

$$CI_{95\%}(Bias_{PMI}) = Bias_{PMI} + / - 1.96\, SE$$

with

$$\begin{aligned} SE &= \sqrt{\frac{1}{f_{A,C}} + \frac{1}{f_{B,C}} + \frac{1}{f_{A,nC}} + \frac{1}{f_{B,nC}}} \\ &\approx \sqrt{\frac{1}{f_{A,C}} + \frac{1}{f_{B,C}}} \end{aligned}$$

This last approximation considers condition 6.

It should be noted that the $Bias_{PMI}$ is not computable if $f_{A,C} = 0$ or $f_{B,C} = 0$. To solve this we use a standard smoothing approach of adding a small value $\epsilon$ to all co-occurrences.

## 3 Experimental setup

To compare the measure of bias based on PMI with those based on embeddings, we follow the approach of Caliskan et al. (2017). They compare biases present in texts with those of two datasets: (1) a 2015 U.S. Bureau of Labor Statistics occupation-gender dataset, which contains the percentage of women in a list of occupations, and (2) a 1990 U.S. census name-gender dataset, which provides, for a list of androgynous names, the percentage of people with each name who are women.

We build a corpus with articles from an English Wikipedia dump from August 2014. We preprocess the dump by removing non alpha-numeric symbols, removing articles with less than 50 tokens and sentence splitting. We use a window size of 10 in all models and we ignore words with less than 100 occurrences, resulting in a vocabulary of 172,748 words.

Word embeddings with 300 dimensions are trained with SGNS and GloVe. For SGNS we use the word2vec implementation of the Gensim library (Řehůřek and Sojka, 2010) with default hyperparameters. GloVe is trained with the original implementation (Pennington et al., 2014) with 100 iterations. This version uses by default additive word representations, in which each word embedding is the sum of its corresponding context and word vectors. For PMI we set the smoothing parameter $\epsilon$ to 0.5.

Using PMI, SGNS and GloVe, we replicate the gender bias experiments performed in the Word Embedding Factual Association Tests (WEFATs) in Caliskan et al. (2017). To represent female and male contexts we use the same lists of words as in Caliskan et al. (2017)[1]. The female proportions for names and occupations in the U.S. were extracted from the datasets provided by Will Lowe's `cbn` R library[2], which contains tools for replicating Caliskan et al. (2017).

For both occupations and names association tests, we assess the correlation between the real-world female proportion and the female bias as measured by PMI, SGNS and GloVe. Female bias refers to the bias metrics where $A$ and $B$ represent the lists of female and male words, respectively. Therefore, positive values imply that the target word is more associated with female words than with male ones.

We emphasize that the objective of this experiment is not to find which method produces a greater correlation between the empirical biases from the U.S. datasets and the textual biases from Wikipedia – the aim is to study whether the three bias metrics produce comparable correlations in these tasks.

## 4 Results

Figure 1 and Table 2 show the scatter-plots and *Pearson's r* coefficients for each of the six experiments (two association tests with three bias measures each). *Weighted Pearson's r* coefficients are shown in Table 2 for the PMI-based metric for both association tests. *Weighted Pearson's r* takes into account the variance of each bias estimate, thus reducing the influence of relatively noisy estimates on the correlation. This type of adjustment is not feasible with embeddings-based methods, which lack a natural notion of statistical variability.

[1] Male terms:*{male, man, boy, brother, he, him, his, son}*. Female terms: *{female, woman, girl, sister, she, her, hers, daughter}*

[2] https://conjugateprior.github.io/cbn/

|  | Occupation-gender association |  |  | Name-gender association |  |  |
|---|---|---|---|---|---|---|
| Metric | SGNS | Glove | PMI | SGNS | Glove | PMI |
| Pearson’s $r$ | 0.70 | 0.70 | 0.69 | 0.77 | 0.74 | 0.78 |
| Weighted Pearson’s $r$ | - | - | 0.79 | - | - | 0.75 |

Table 2: *Pearson’s r* coefficients for each experiment shown in Figure 1. Because the PMI-based metric provides a measure of variance, *weighted Pearson’s r* coefficients have been calculated to account for the variability of each data point.

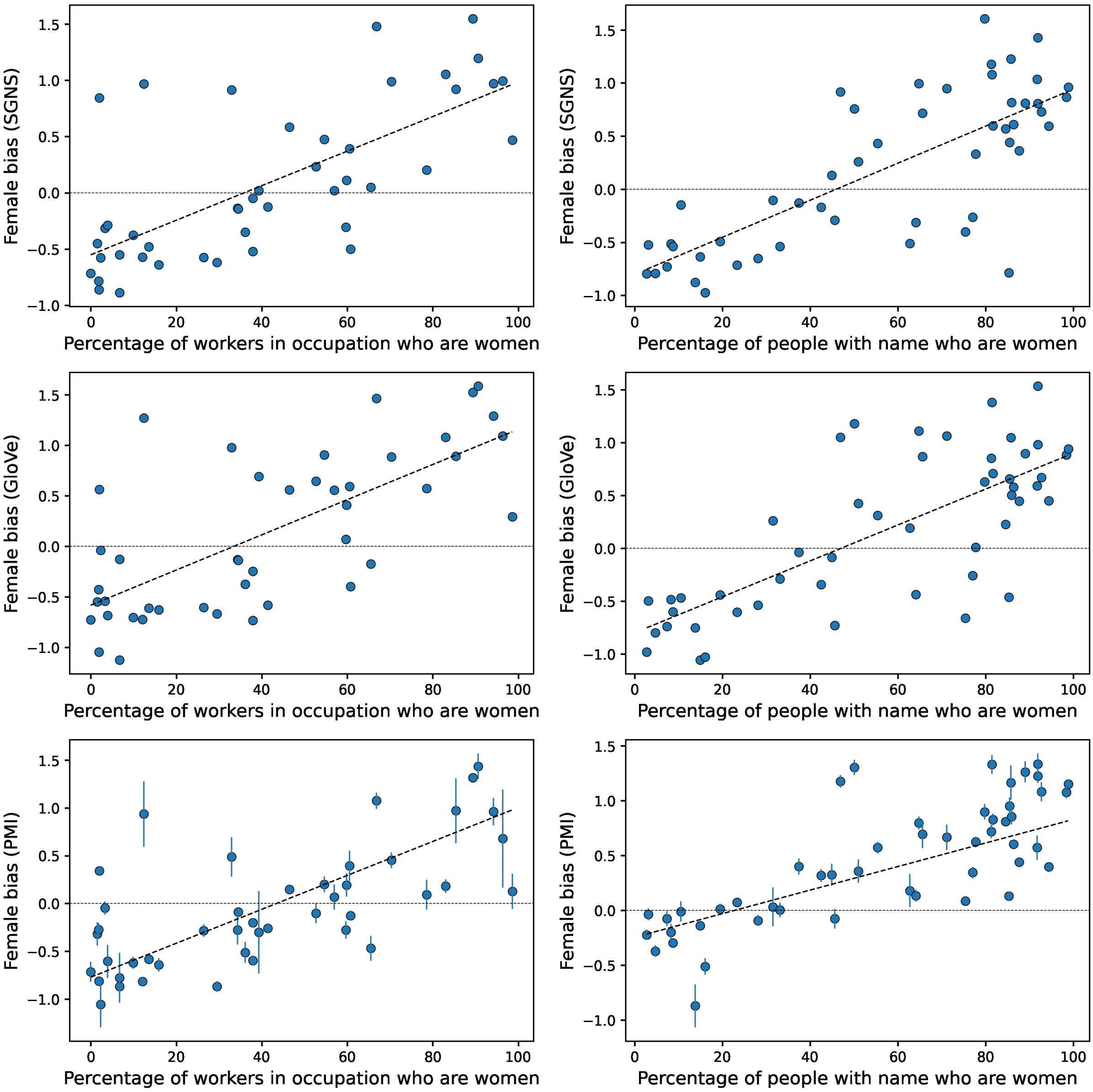

Figure 1: Occupation-gender association experiment (left panels) and names-gender association experiment (right panel) for female bias measures based on SGNS (top panels), GloVe (middle panels), and PMI (bottom panels). In dashed lines linear regressions are shown – in the case of PMI-based bias the variability was taken into account as weights. Error bars in PMI-based measure represent confidence intervals.

In the androgynous names setup, we find similar degrees of linear correlation for the three bias measurement methods, for both weighted and unweighted correlations. In the occupations experiment, we find similar values of *Pearson’s r* for the three methods; and an increase in the correlation is observed when the variability of bias is considered in the estimation.

All in all, the embeddings-based methods and the PMI-based method have comparable performances in the WEFAT tasks of Caliskan et al. (2017).

### 4.1 Significance testing

A permutations test that shuffles context words is a way that has been used in embeddings-based bias metrics in order to calculate statistical significance (Caliskan et al., 2017; Garg et al., 2018; Charlesworth et al., 2021).

A limitation of this technique is that, for the test to have enough power, many words are required in the lists. To show this, we perform permutation tests in the occupations experiment using the SGNS-based method. We shuffle the words from lists $A$ and $B$ repeatedly and estimate the probability of obtaining an absolute value of bias that is equal to or greater than the one obtained in the original configuration (North et al., 2002).

In Figure 2 we compare the p-values of the permutation test in the SGNS-based metric with the p-values of the log odds ratio test of the PMI-based metric. A Benjamini-Hochberg correction was applied to the p-values obtained by both methods to account for multiple comparisons.

The permutation test indicates that only the two words with the highest female SGNS-based bias are significantly different from zero at a 0.05 significance level. This supports our hypothesis that permutations tests don't have enough power whenever word lists are small. In contrast, in the case of the PMI-based metric, the log odds ratio test indicates that the majority of points are significantly different from zero, with the exception of some points with bias values close to zero.

## 5 Interpretability

Model interpretability has become a core topic of research in NLP. Loosely speaking, it refers to the degree to which a human can understand the cause of a decision (Miller, 2019) – which has become progressively more complex as current models steer away from simpler setups.

Although there are many studies on how the vector space of word embeddings is formed (Levy and Goldberg, 2014; Levy et al., 2015; Ethayarajh et al., 2019), there is no transparent interpretation of bias measurements formed by cosine distances between word vectors. This can be partially observed in the growing literature which tries to construct metrics in order to measure word embedding interpretability (Şenel et al., 2018; Jang and Myaeng, 2017; Panigrahi et al., 2019). Most of those metrics can be considered *post-hoc* [3].

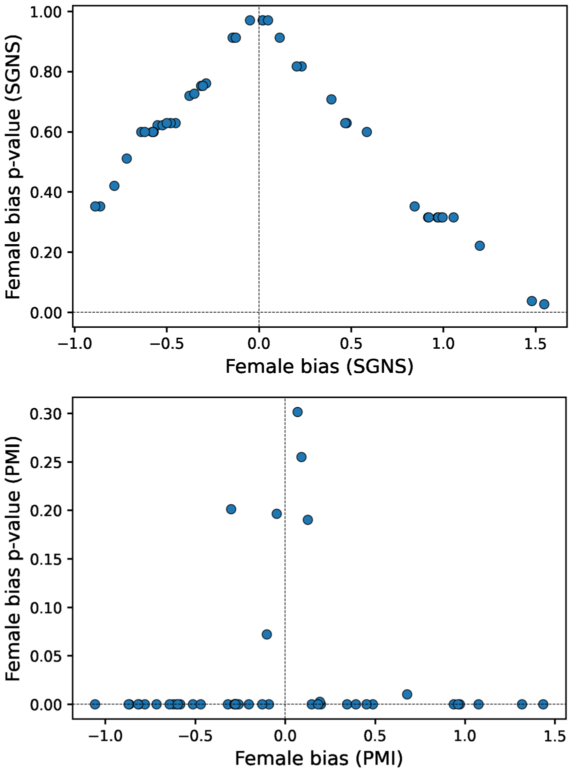

Figure 2: Statistical significance of the permutations test in the SGNS-based metric vs. the SGNS-based bias value (top panel), and significance of the odds ratio test vs. the PMI-based bias value.

In contrast, the PMI-based bias measure can be expressed intrinsically in terms of conditional probabilities:

$$Bias_{PMI}(C) = log\left(\frac{p(C|A)}{p(C|B)}\right)$$

The bias is interpreted as the logarithm of how many more times it is likely to find $C$ in the context of words in $A$ than in the context of words in $B$.

For example, in the Wikipedia corpus the $Bias_{PMI}$ of word *nurse* is 1.3159, thus,

$$\frac{p(nurse|A)}{p(nurse|B)} = exp(1.3172) = 3.7330$$

This can be interpreted as stating that it is 273.30% more likely to find the word *nurse* in the context of words associated with women ($A$) than in the context of words associated with men ($B$).

[3] Rudin (2019) makes a case for designing intrinsically interpretable models instead of *post-hoc* evaluated models.

## 6 Discussion

As far as we know, our bias metric is the first that provides a simple and efficient way of evaluating the statistical significance of the obtained biases.

We highlight the importance of knowing if the measured patterns are indeed significant. Imagine the case where we want to know if a collection of texts has a particular bias. To address this situation, it is not only necessary to estimate the magnitude of the bias, but also to have a decision rule that lets us know up to what degree this value might have been due to statistical fluctuation. A statistical test associated with the bias measure is essential in this use case.

SGNS and GloVe embeddings can capture word associations of second order or higher (Levy et al., 2015; Altszyler et al., 2018; Schlechtweg et al., 2019). Therefore, when embeddings are used to measure word associations, it is not possible to know whether the emerging biases are due to ostensive first-order co-occurrences or whether they are derived from higher-order co-occurrences. Rekabsaz et al. (2021) evidenced the presence of these spurious associations when quantifying semantic similarity with second-order contributions, and Brunet et al. (2019) showed in a sensitivity analysis that second-order effects have an important contribution in the quantification of bias with the WEAT metric from Caliskan et al. (2017). Nevertheless, bias metrics based on second-order associations have the advantage of managing data sparsity. For example, they can capture relevant associations in synonyms occurrences of target and context words. In the case of our first-order metric, this problem must be addressed by increasing word lists with synonyms and different forms of the words of interest.

Rekabsaz et al. (2021) propose a metric based on word embeddings that only captures first-order associations. Their results show that their metric correlates with biases in employment more than other measures based on second-order co-occurrences. This measure uses the product of word and context matrices of the SGNS model (usually known as $W$ and $C$), so this measure can be interpreted as a shifted-PMI with smoothing. In future work we will compare Rekabsaz et al. (2021) metric with ours. However, we emphasize that the benefit of using the PMI-based metric lies in the interpretability and the estimation of confidence intervals and statistical significance.

## 7 Conclusions

In this article we present a PMI-based metric to capture biases in texts, which has the benefits of (a) providing simple and computationally inexpensive statistical significance tests, (b) having a simple interpretation in terms of word co-occurrences, and (c) being explicit and transparent in the associations that it is quantifying, since it captures exclusively first-order co-occurrences. We show our PMI-based bias measurement can be approximated by a log odds ratio. This allows for the calculation of confidence intervals and statistical significance for bias, using as the null hypothesis the absence of bias. We demonstrate that our measure can be expressed as the logarithm of the ratio of the probabilities of the target words conditional on the context words. This provides a simple interpretation of the metric's magnitudes in terms of word co-occurrences: how much more likely is it to find target words $C$ in a window around context words $A$ than in a window around context words $B$?

Finally, we prove that the use of our method produces a performance comparable to the one produced by GloVe-based and Skip-gram-based metrics in Caliskan et al. (2017)'s experiments of gender-occupation and gender-name associations.

## References

Alan Agresti. 2003. *Categorical data analysis*, volume 482. John Wiley & Sons.

Edgar Altszyler, Mariano Sigman, and Diego Fernández Slezak. 2018. Corpus specificity in LSA and word2vec: The role of out-of-domain documents. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

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*, volume 29. Curran Associates, Inc.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

David DeFranza, Himanshu Mishra, and Arul Mishra. 2020. How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*, 119(1):7–22.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Robert M Fano. 1961. *Transmission of Information: A Statistical Theory of Communication*. MIT Press.

Ramiro H. Gálvez, Valeria Tiffenberg, and Edgar Altszyler. 2019. Half a century of stereotyping associations between gender and intellectual ability in films. *Sex Roles*, 81(9):643–654.

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):E3635–E3644.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Kyoung-Rok Jang and Sung-Hyon Myaeng. 2017. Elucidating conceptual properties from word embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 91–95, Online. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Alison P. Lenton, Constantine Sedikides, and Martin Bruder. 2009. A latent semantic analysis of gender stereotype-consistency and narrowness in american english. *Sex Roles*, 60(3):269–278.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4(10):1021–1028.

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*, volume 26. Curran Associates, Inc.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

B. V. North, D. Curtis, and P. C. Sham. 2002. A note on the calculation of empirical p values from monte carlo procedures. *The American Journal of Human Genetics*, 71(2):439–441.

Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. 2019. Word2sense: sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5692–5705, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Navid Rekabsaz, Robert West, James Henderson, and Allan Hanbury. 2021. Measuring societal biases from text corpora with smoothed first-order co-occurrence. *Computing Research Repository*, arXiv:1812.10424.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Dominik Schlechtweg, Cennet Oguz, and Sabine Schulte im Walde. 2019. Second-order co-occurrence sensitivity of skip-gram with negative sampling. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 24–30, Florence, Italy. Association for Computational Linguistics.

Lütfi Kerem Şenel, İhsan Utlu, Veysel Yücesoy, Aykut Koç, and Tolga Çukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779.