Abstract

In this report, the current literature surrounding bias in pre-trained word embeddings is examined. Due to a lack of interpretability, flexibility, and concerns for transparency over existing measures of bias, a new method for identifying and measuring biases is described. Using three pre-trained word embeddings, we investigate the types and levels of racial bias found in educational writing, news, and social media platforms. A level of statistical significance was reported with each result using a permutation test. Using the method described in this report, no significant explicit racial biases were found. However, significant implicit biases were found to exist in all three word-embeddings. This demonstrates the ability for word-embeddings to contain human-like biases. The failure to detect explicit racial bias signals the decline in overt racism. A unanimous detection of implicit bias using stereotypical names for each race, highlights the failure of society to free ourselves from the shackles of racism and racial bias.

1 Problem

Racism and racial bias has plagued humanity for hundreds of years. This has lead to racism, and thus racial bias, being ingrained into the English language. While racial bias may exist towards many ethnicities and minorities, we shall focus on racial bias towards people who identify as black or of African descent. More importantly, we shall examine the difference between this ethnic group and the group which forms the ethnic majority in the English language; white or Caucasian[1]. The reason for this choice is the existing relationship between negative connotation and the word black; black-hearted, black-mail, or black-list. Conversely, there exists a positive connotation with the word white; white-knight, white-lie, or white-hope. Therefore, there is a strong case that racial bias towards black people has been explicitly ingrained in the English language. However, as overt racism declines these words have become less prominent This does not mean that racial bias has disappeared from our language. One study found that African-American names were more likely to be associated with unpleasant terms and European-American names are more likely to be associated with pleasant terms [1]. We shall attempt to compare the prominence of racial bias present in two pre-trained GloVe models [2] and one pre-trained word2vec model[3]. Borrowing terminology from previous work in the area, this analysis will require a set of attribute words to be defined. These will capture the different attributes of each race. A set of target words, which should be neutral in a racial context, are used to identify bias. The pre-trained models are trained on data from Wikipedia, Google News, and Twitter [4] and will have 300, 300, and 200 dimensions respectively. The aim of this analysis is to identify the different types and levels of racial bias found in educational writing, news, and social media platforms.

2 Literature Review

Natural Language Processing has been an emerging field over the last decade. It has many applications from 'smart' assistants like Apple's Siri, to spam filtering and CV parsing. Much of the early work in this area relied on occurrence statistics. These were simple models which failed to capture subtle contextual meanings. These types of models would be unsuitable for the proposed application in this report.

The word2vec model is trained on the Google News data-set. This corpus is by far the largest of the three seen with 1.6 billion words. The word2vec model was designed and patented by Google to allow for the training of models on large corpora. It uses a shallow feed forward neural network, with two potential architectures. The first of which is known as a Bag-of-words and the second is known as Skip-Gram. As this model is pre-trained and its training is not the focus of this review, further details

[1]This misuse of Caucasian is to reflect its common usage in the USA, where Caucasian was adopted as a synonym for white. Its original meaning was to indicate a person who originated from the Caucasus

can be found in the previous references. One of the key points of the word2vec model developed by Google is the lack of occurrence statistics and the use of word analogy tests to evaluate the model. Word analogies allow a more comprehensive overview of the overall semantics within the model. This makes it highly appropriate when investigating subtle contextual biases such as gender or race.

The Wikipedia and Twitter trained models considered in this paper are referred to as Global Vector (GloVe) models. The GloVe method is an unsupervised learning method which aims to encapsulate the semantic meanings of words into its vector representation. One primary focus of the GloVe model is to represent a "global" view of the corpus of interest. The model attempts to summarise subtle contextual relationships better than previous models through occurrence and more importantly, co-occurrence statistics and probabilities. This was a key point for the GloVe model as it aimed to further the improvements of Google's word2vec model, without losing the reliance on occurrence statistics. GloVe focuses on ratios of co-occurrences and utilises this information in a log bi-linear regression model. The GloVe model also used word analogies as one of the primary methods of evaluation. Similarly, as the training of these models is not the primary focus of this review, more information can be found in the previous references.

Bias contained in word embeddings has become a new area of research in recent years due to their prevalent use in many applications. For example, gender bias was shown to exist in a coreference resolution system due to bias in the word embedding used [5]. With the development of new, advanced, contextual-based methods such as GloVe, subtle biases are being encoded into these word embeddings. This has lead to de-biasing algorithms being proposed to remove bias from corpora and allow for unbiased word embedding. Many de-biasing examples focus on identifying and removing gender bias [6]. We are only going to discuss the identification process used, as the de-biasing algorithm is not of interest. It is difficult to pin-point the reason that gender bias has been the focus of many de-biasing algorithms, however it may be to allow easy comparison to other proposed methods or due to the over reliance on antithetical pairs. The method used to identify bias in [6] relies on constructing a gender sub-space, $g \in \mathbb{R}^d$, using Principal Component Analysis (PCA). This is done by aggregating multiple antithetical pairs, for example {he,she}, {man,woman} etc. Using multiple pairs increases robustness to polysemy, i.e. when words have multiple, unrelated meanings. These antithetical pairs are converted to their vector representation and then the vector difference is calculated. The resulting vectors are then reduced to the single vector g, which represents a gender subspace/direction. A set of target words, N, which should be gender neutral are identified and the "direct bias" is calculated using the following formula:

$$\frac{1}{|N|}\sum_{w\in N}|\cos(w,g)|^c$$

This formula has a parameter c which controls the "strictness" of the measurement, with 0 being the most strict. This parameter, despite providing a degree of control to the user, allows for subjective tuning of the bias detected. This may be necessary when focusing on removing bias, however, it is inappropriate to allow this subjectivity when trying to identify if bias exists. It is possible for this parameter to be tuned with the aim of maximising the bias detected and inflating results artificially. Another problem with this method is the apparent lack of scale. The paper for example suggests that a direct bias value of 0.08 suggests "that many occupation words have substantial component along the gender direction." This is not apparent, as the mean of the absolute cosine similarity raised to some power c is not a known scale. This paper also utilised many "crowd-workers" to evaluate whether analogies displayed gender bias or not. There is little information of the composition of these crowd-workers to ensure the judgments were objective or unbiased. While this paper forms a strong basis of de-biasing methods, the methods of quantifying bias are not generally applicable.

Many psychology papers investigate bias in language and culture. This is usually done with questionnaires and small sample sizes. These surveys can be poorly constructed and the small sample sizes result in the statistical conclusions having a low power. In particular, a paper from the University

of Washington found racial bias to exist by using a set of typical European-American names and a set of typical African-American names. These were used in conjunction with a set of target words which represented pleasant and unpleasant sentiments. This paper used an implicit association test to show that the African American names were more likely to be associated with unpleasant sentiment [7]. The set of names examined were chosen by a group of introductory psychology students as being more likely to belong to one group when compared to the other. One paper successfully replicates the findings of racial bias using the same sets of attribute and target words on the pre-trained "Common crawl" GloVe model [1]. No direct level of bias is reported in the paper, however an effect size and a permutation test p-value are reported. The effect sizes are large (1.28-1.5), and the p-values are small $(10^{-8} - 10^{-3})$. However, the effect sizes are unbounded above, and while a general rule of thumb exists for interpreting effect sizes, they are not a concrete measure of bias. Thus, this measure of bias is incomplete. It should also be noted that the method used to select the names is subjective. These names were chosen as they were associated with one race more than the other and therefore are loaded with bias. This is perfect in the application of the paper, which was identifying if the GloVe model contained human like biases. However, they are less suitable for identifying if racial bias exists in the corpus outright. The permutation test used permutes attribute words across sets and calculates the difference in the mean cosine similarities of target and attribute words. A permutation test is an appropriate method of testing the significance of the bias.

Therefore, in the current literature there is a lack of a definitive measure of bias. There is also a lack of objectivity when defining the bias or constructing tests for bias. This can be greatly improved upon by being more transparent with the processes used to classify and investigate bias.

3 Plan

To compare levels of racial bias in the different models, we are required to define bias mathematically. One option is to adapt the method seen in [6], and construct a racial sub-space using PCA. This approach requires the construction of a set of antithetical attribute pairs. The many pair-wise antithetical words required are not as readily available as in the case of gender. The requirement for multiple pairs is to increase robustness to polysemy and random variations due to word count in a corpus. For example, the initial pairing of black and white encounters the first problem, as these refer to generic colours which can be used in many different settings and not just when referring to race. Therefore, multiple pairings are necessary. One alternative method is seen in the documentation for the debiaswe package on Github[2]. This package was a direct result of the paper referenced above. The lack of antithetical pairs can be addressed by adapting this method. The summed vector of each attribute set is calculated and then the vector difference is calculated. This has two advantages; first, the need for pairs across attribute sets is removed, and second, the racial subspace obtained is always a single p-dimensional vector, where p is the dimensions of the word embedding. This means the need for PCA is also removed. We shall make a further adaptation to this method. As the method relies on the summed vector, we are required to have equal cardinality for both attribute sets. This may be a logical requirement, however in some circumstances it may not be possible. Using the mean vector instead of the summed vector will allow for unequal cardinality of sets. In the case where the cardinality is equal, the mean vectors will only differ in magnitude from the summed vectors, meaning the racial direction is preserved. This allows for the specification of the attribute set to be more flexible, which will allow for easier use investigating more complex bias in the future.

Bias in this case could be considered the direction and magnitude of the projection of the set of target words onto the racial subspace, i.e. projecting the target words on to the difference vector described above, could represent the bias of that word. As we are only aiming to investigate racial bias

[2]This can be found at: https://github.com/tolga-b/debiaswe

in a broader perspective and are not required to de-bias the model, it is not required to classify the bias of each word. If it were possible to establish a semantic subspace/direction from opposing sets of target words, this could allow a comprehensive measure of bias to be established. Using a similar method to the above would require the specification of a set of negative and a set of positive target words. This adds a layer of subjectivity, as the target words must be classified as either positive or negative. This replicates the approach taken in [1] by classifying words as pleasant and unpleasant. The hypothesis would be that this semantic sub-space/direction should be orthogonal to the racial sub-space should no racial bias exist. Therefore, as both subspaces would be represented by a single column vector, the inner product between vectors provides a measure of bias which is readily interpretable. An inner product with absolute value close to 0, will suggest that no bias exists between the vectors as they are almost orthogonal. An inner product with absolute value close to 1, will suggest that a large amount of bias exists, with the racial direction being almost parallel to the semantic direction. The sign of the inner product will dictate the direction of the bias. Comparisons between models may be misleading as the inner product is not a linear scale between -1 and 1. It may be preferable for some to report the arc-cos of the inner product, this will convert the bias to a linear scale. A permutation test can be used to identify the significance of the bias. This leads to the following steps:

1. Specify a set of attribute words related to white and black people denoted A_1 and A_2.
2. Calculate the mean vector representation of these sets, $\bar{a}_1$ and $\bar{a}_2$, $\quad \bar{a}_1, \bar{a}_2 \in \mathbb{R}^d$.
3. Define the racial subspace/direction, $r = \bar{a}_1 - \bar{a}_2, \quad r \in \mathbb{R}^d$
4. Specify a set of positive semantic target words and negative semantic target words denoted T_1 and T_2.
5. Calculate the mean vector representation of these sets, $\bar{t}_1$ and $\bar{t}_2$, $\quad \bar{t}_1, \bar{t}_2 \in \mathbb{R}^d$.
6. Define the semantic subspace/direction, $s = \bar{t}_1 - \bar{t}_2, \quad s \in \mathbb{R}^d$
7. Calculate the inner product as a measure of bias, a positive inner product will represent a positive bias towards white people. A negative inner product will represent a positive bias towards black people.
8. Conduct a permutation test by permuting the attribute words between A_1 and A_2. This will allow for it to be determined if the level of bias found is statistically significant.

The null hypothesis of the permutation test is that the distribution of the vectors in the attribute sets A_1 and A_2 are identical. This hypothesis can be tested by randomly permuting the elements between these attribute set and recomputing the bias according to the above algorithm. In a full permutation test, the bias of all permutations are computed.This would require $|A_1| + |A_2|!$ computations. Therefore, a sample of 10,000 permutations are used and a confidence interval for the generated p-value shall be supplied. If either bound of this confidence interval is greater than the power of the test, e.g. 0.05, then we shall fail to reject the null hypothesis.

As the algorithm for calculating bias in a corpus has been defined, we shall now define the attribute and target sets that we shall be testing on the three pre-trained word embeddings. We shall test each model for three different types of bias. These will include explicit bias by defining a set of attribute words which explicitly describe the differences between the races. Stereotypical implicit bias shall be tested by using the set of stereotypical names defined in [1]. General implicit bias shall be tested by using a combination of the most common baby names by race. For the latter, the data used is the 25 most common baby names by gender and mother's ethnicity in the City of New York from 2011 to 2016 [3]. Naming patterns in New York City may not be representative of the whole of the US. The data was used as it was easily accessible and of open record. Thus, it provided an objective manner in which to identify names for use. The full details of each set of attribute words is available in Appendix B.

We are also required to define positive and negative semantic target sets. Selecting general adjectives

[3]This can be found at: https://www.kaggle.com/new-york-city/nyc-most-popular-baby-names

with positive and negative semantics can be difficult due to the subjectivity of positive and negative. It is possible that the choice of these words could be manipulated to ensure bias was found. It also should be noted that choosing a wide range of topics or mis-classifying positive or negative semantics may disguise biases. If the vector representation of the elements of a target set are evenly distributed in the racial direction, the resulting mean vector shall be orthogonal relative to r. This is a large flaw in the current design of this measure of bias, as correctly identifying positive and negative semantics may require many crowd-workers or may encapsulate a user's personal biases. For example, some people may consider a skilled-craftsman a negative target word when compared to an office worker. Due to these reasons, for this investigation a decision to focus on topics relating to intelligence, education, and employment was made. Intelligence, well-educated, and skilled/trained employment were considered the positive semantic. The words chosen were explicitly linked to one of these topics and many appropriate words were not included as the size of the target sets were small in an attempt to avoid personal bias. A full list of the target words used are available in the appendix.

4 Results & Discussion

Given the three pre-trained word-embeddings, three combinations of attribute sets; Explicit, Implicit, and Stereotypical, and one combination of target sets, there is a total of nine measures of bias. The values of these are reported below along with a p-value from a permutation test and the 95% confidence interval for the p-values.

Model (Attribute sets)	Bias	p-value	p-value 95% CI
Google (Implicit)	0.0752	0.0748	(0.0696, 0.08)
Google (Explicit)	0.1242	0.1432	(0.1363, 0.1501)
Google (Stereotypical)	0.1848	0.008	(0.0063, 0.0097)
Wikipedia (Implicit)	0.3106	0.0034	(0.0023, 0.0045)
Wikipedia (Explicit)	0.0839	0.227	(0.2188, 0.2352)
Wikipedia (Stereotypical)	0.3502	3e-04	(0, 6e-04)
Twitter (Implicit)	0.0879	0.1745	(0.1671, 0.1819)
Twitter (Explicit)	0.0985	0.2292	(0.221, 0.2374)
Twitter (Stereotypical)	0.1407	0.0457	(0.0416, 0.0498)

Table 1: Bias and significance

As can be seen in table 1, four models display a significant positive racial bias towards white people at the 5% level of statistical significance. The other five models also display a positive racial bias towards white people, however it is not significant at the 5% level. None of the models display bias when using the explicit racial attribute sets and all of the models display a racial bias using the stereotypical attribute sets. Only Wikipedia displays a significant racial bias when using the implicit attribute sets.

The lack of significant racial bias when using the explicit attribute sets may support the decline in overt racism in modern society. However, it may also be due to a poorly constructed attribute set. Both attribute sets used contain four words each. This is a much lower figure than the other combinations of attribute sets. This means that the decision to include "negro" in the attribute set associated with black people may have been sub-optimal, as this word has fallen out of use in most forms of language. As it was the only racist word included in any of the attribute sets, it may have caused the mean vector to change drastically. This could be investigated by comparing the difference between the mean vector of the attribute set with, and without its inclusion. The decision to include it was taken in an attempt to capture a more direct version of racial bias.

The stereotypical attribute sets produced the most interesting results. Despite all models finding a statistically significant bias, the magnitude of that bias differs greatly. The explicit bias found for the Wikipedia trained model is the highest of any model tested at 0.3502. This was almost double the amount of bias found for the Google model and close to 2.5 times when compared to the Twitter trained model. However, recall that the inner product is not a linear scale and so converting these values to degrees, we see that Wikipedia has an angle of 70 degrees compared to Twitters angle of 80 degrees. This is still a notable difference in the levels of bias.

The implicit bias attribute sets offer a unique insight into the workings of the algorithm. In order to visualise the algorithm and the results seen, we can project each word in the positive and negative target sets onto the racial subspace vector.

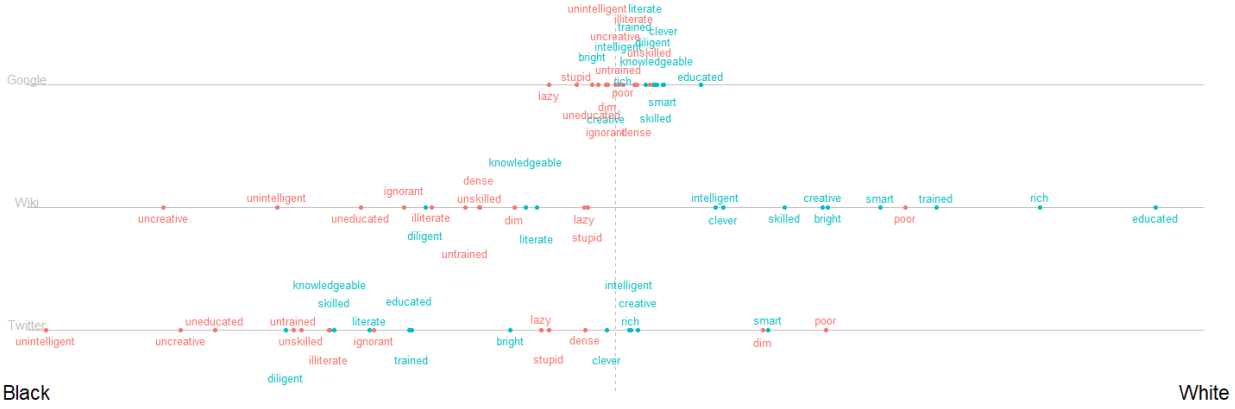

Figure 1: Visualisation the target word sets on the Implicit Bias subspace vectors

Note: words are jittered in the vertical axis to avoid overlap and this axis contains no meaning. The x-axis, which represents the racial subspace vector, is linearly scaled. The dotted line represents the point at which target words are race neutral.

Looking at Fig. 1, we can see the results of such a visualisation. This visualisation provides great insight into the individual measures of bias and the p-values associated with them. Examining the Google subspace, we note that most of the words, both positive and negative, are near the race neutral line. However, there appears to be a slight leaning to a positive bias towards white people as there are only 2-3 positive words to the left of the dotted line. This agrees with the low measure of bias computed. The p-value is insignificant but is relatively near the level of significance. It may be concluded that bias would be present with a different selection of target words, albeit at a low level. The Wikipedia model has the second largest bias of all models. This is intuitive as the majority of the positive words lie to the right of the race neutral line, with the majority of negative words to the left. We examine the positive word most associated with the black attribute set, diligent. This may represent diligence being a trait which is associated with physical work or less skilled work as it appears in this position for multiple models. The final model, Twitter, has a similar level of racial bias when compared to the Google model. However, the p-value is much larger. It appears for this model that most of the target words occur to the left the race neutral line. They are also relatively evenly dispersed, which explains the near-orthogonal inner product. We can see three very different levels of bias present, and this visualisation provides some potential explanation for these differences.

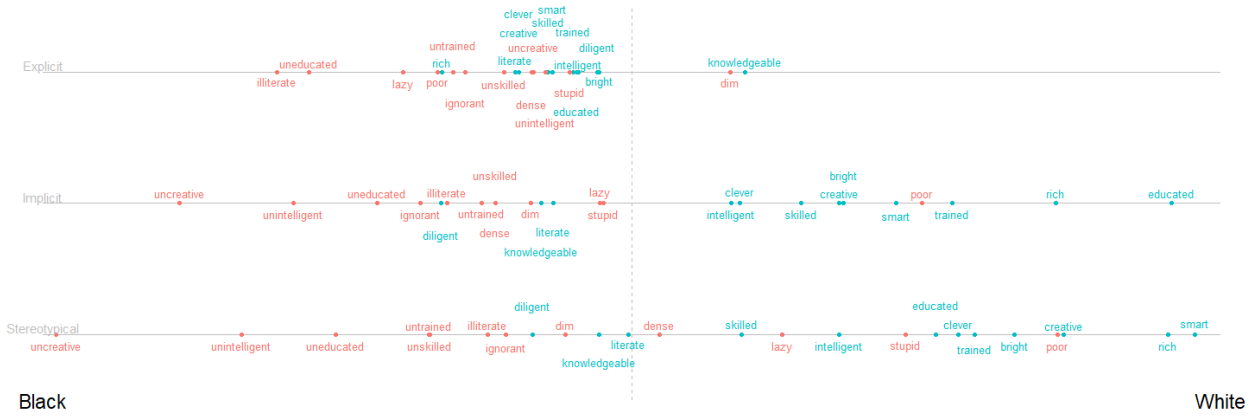

Figure 2: Visualisation the target word sets on the Wikipedia trained model

Looking at Fig.2, the Wikipedia model offers some interesting insights of its own. For the explicit attribute set, we see a similar pattern to Twitter in Fig.1. Most of the words, both positive and negative appear to the left of the race neutral line. This appears to correlate to the level of bias found and the associated p-value which is also similar to that of the Twitter model. The stereotypical attribute axis, also has similar properties to the implicit axis which was seen in Fig.1. The majority of the positive target words lie to the right of the race neutral line. This affirms the very similar levels of bias and significance seen for both attribute sets. This visualisation gives a comprehensible view of the high dimensional space that the racial and semantic subspace vectors are defined in. They also suggest that the intuitive information contained in them can be represented by a single and comparable measure of bias, provided they are accompanied with an appropriate level of significance.

5 Conclusion

The current literature surrounding the identification of human-like biases in word embeddings is sparse. Many methods focus on removing bias once identified, rather than developing a consistent, comparable, and flexible method for the identification process. The examples seen in Section 2 of this report highlight the need for a comparable measure of bias which is on a consistent and bounded scale. The method proposed in this project is a known quantity, and can be easily converted to a linear scale if required. It allows for easy comparison between models and across word embeddings. It is flexible in the specification and cardinality of attribute and target sets. Though, as with all methods, particular care is needed with the specification of these sets. As seen in the visualisations presented in Fig. 1 and Fig. 2, these sets can be altered to manipulate the level of bias found or its statistical significance. Therefore, it is required that the method used for constructing these sets is as objective and transparent as possible. The permutation test suggested in this report allows for a level of statistical significance to be associated with the measure of bias. This level of statistical significance appears to correspond to the geometric interpretation, i.e. bias which is statistically insignificant appears to be in agreement with geometry presented in the visualisations. The simplicity of this measure of bias also lends itself to the visualisation process.

Using the method outlined, three types of racial bias were investigated in three separate pre-trained word embeddings. Of the nine combinations examined, four were found to contain statistically significant levels of bias. However, these findings were subjected to a range of decisions regarding the choice attribute and target sets. Explicit bias was found to exist in none of the three models. This suggests the explicit racism towards people of African descent has finally begun to disappear from society and language. Implicit bias was found to be statistically significant only in the case of the

Wikipedia model. It is difficult to draw conclusions from this result in isolation, as the attribute sets were constructed using data which may not be representative of the US. If this result is considered in conjunction with the result seen with the stereotypical attribute set, it may be concluded that the implicit bias is due to history related Wikipedia entries. Historical biases may have lead to people of African-American descent being under-represented on Wikipedia. It may be argued that the levels of implicit bias seen in the Google and Twitter embeddings are low. It should be noted that these results were shown to be statistically significant at the 5% level.

References

[1] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[2] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[4] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010. http://is.muni.cz/publication/884893/en.

[5] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint arXiv:1804.06876*, 2018.

[6] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), pp. 4349–4357, Curran Associates, Inc., 2016.

[7] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz, "Measuring individual differences in implicit cognition: the implicit association test.," *Journal of personality and social psychology*, vol. 74, no. 6, p. 1464, 1998.

A Reproducible example in R

```
library(reticulate)

use_python("/usr/bin/python3")
gensim <- import("gensim")

get_vector <- function(model, word){model$get_vector(word)}
vget_vector <- Vectorize(get_vector, "word")

inner_product <- function(a,b){sum(a*b) / ( sqrt(sum(a * a)) * sqrt(sum(b * b)) )}

create_subspace <- function(model, words_1, words_2){
  vectors_1 <- vget_vector(model, words_1)
```

```
  vectors_2 <- vget_vector(model, words_2)
  mean_vector_1 <- apply(vectors_1, 1, mean)
  mean_vector_2 <- apply(vectors_2, 1, mean)
  difference <- mean_vector_1 - mean_vector_2
  list("Subspace" = difference, "Vectors_1" = vectors_1, "Vectors_2" = vectors_2)}

permutation_test <- function(N, current_bias, model, attr_vectors_1, attr_vectors_2, tar_
    ↪ subspace){
  results <- rep(F,N)
  l1 <- ncol(attr_vectors_1)
  l2 <- ncol(attr_vectors_2)
  attr_set <- matrix(c(attr_vectors_1, attr_vectors_2), ncol = l1+l2)
  for (i in 1:N){
    perm <- attr_set[,sample(1:(l1+l2))]
    mean_vector_1 <- apply(perm[,1:l1], 1, mean)
    mean_vector_2 <- apply(perm[,(l1+1):(l1+l2)], 1, mean)
    perm_difference <- mean_vector_1 - mean_vector_2
    result <- inner_product(perm_difference, tar_subspace)
    results[i] <- (result >= current_bias)}
  pval <- mean(results)
  CI <- pval + sd(results)/sqrt(length(results))*qnorm(c(0.025,0.975))
  list("P_value" = pval, "CI" = CI)}

bias <- function(model, attr_words_1, attr_words_2, tar_words_1, tar_words_2){
  attribute_subspace <- create_subspace(model, attr_words_1, attr_words_2)
  target_subspace <- create_subspace(model, tar_words_1, tar_words_2)
  bias <- inner_product(attribute_subspace$Subspace, target_subspace$Subspace)
  p_val <- permutation_test(10000, bias, model, attribute_subspace$Vectors_1, attribute_subspace
      ↪ $Vectors_2, target_subspace$Subspace)
  list("Bias" = bias, "P_value" = p_val$P_value,"95_CI" = p_val$CI, "Attribute_subspace" =
      ↪ attribute_subspace, "Target_subspace" = target_subspace)}

wiki <- gensim$downloader$load('glove-wiki-gigaword-300')

white_names <- c('Adam','Chip','Harry','Josh','Roger','Alan','Frank','Ian','Justin','Ryan', 'Andrew
    ↪ ','Fred','Jack','Matthew', 'Stephen','Brad','Greg','Jed','Paul','Todd','Brandon','Hank','
    ↪ Jonathan','Peter','Wilbur','Amanda','Courtney','Heather','Melanie','Sara','Amber','Crystal','
    ↪ Katie','Meredith','Shannon','Betsy','Donna','Kristin','Nancy','Stephanie','Ellen','Lauren','
    ↪ Peggy','Colleen','Emily','Megan','Rachel','Wendy')
black_names <- c('Alonzo','Jamel','Lerone','Theo','Alphonse','Jerome','Leroy','Torrance','Darnell','
    ↪ Lamar','Lionel','Rashaun','Tyree','Deion','Lamont','Malik','Terrence','Tyrone','Lavon','
    ↪ Marcellus','Nichelle','Shereen','Ebony','Latisha','Shaniqua', 'Jasmine','Latonya','Tanisha','Tia
    ↪ ','Lakisha','Latoya','Yolanda','Malika','Tawanda','Yvette')

positive_word_list <- c("intelligent", "bright", "smart", "clever", "educated", "knowledgeable", "
    ↪ literate", "trained", "skilled", "creative", "rich", "diligent")
negative_word_list <- c("unintelligent", "dim", "stupid","dense", "uneducated", "ignorant", "
    ↪ illiterate", "untrained", "unskilled", "uncreative", "poor", "lazy")
set.seed(0)
```

```
wiki_results_stereo <− bias(wiki, tolower(white_names), tolower(black_names), positive_word_list,
    ↪ negative_word_list)
```

B Attribute and Target sets

B.1 Implicit

A_1 = {David, Joseph, Moshe, Jacob, Michael, Benjamin, James, Daniel, Alexander, Jack, Samuel, John, Adam, Matthew, Henry, Chaim, Abraham, Nicholas, Ryan, William, Ethan, Liam, Noah, Charles, Thomas, Esther, Leah, Sarah, Olivia, Chaya, Rachel, Emma, Ava, Miriam, Sophia, Emily, Ella, Chana, Mia, Isabella, Charlotte, Sofia, Maya, Rivka, Sara, Alexandra, Abigail, Elizabeth, Anna, Victoria}

A_2 = {Noah, Elijah, Aiden, Jeremiah, Jayden, Ethan, Josiah, Joshua, Amir, Mason, Tyler, Liam, Christian, Michael, Isaiah, Jacob, Carter, Ayden, Justin, David, Christopher, Chase, Daniel, Malachi, Logan, Madison, Ava, Aaliyah, Chloe, London, Taylor, Kayla, Olivia, Nevaeh, Serenity, Skylar, Fatoumata, Abigail, Savannah, Gabrielle, Brielle, Arianna, Nyla, Faith, Khloe, Ariel, Isabella, Mia, Mariam, Leah }

Note: Mikayla was replaced in A_2 with the 26th most common name, Leah, due to Mikayla not occurring in the Wikipedia corpus. Also, the intersections between A_1 and A_2 were removed as their removal does not effect the mean vector of either set. It also stopped the intersection interfering with the permutation test.

B.2 Explicit

A_1 = {white, caucasian, european, light}

A_2 = {black, african, dark, negro}

B.3 Stereotypical

A_1 = {Adam, Chip, Harry, Josh, Roger, Alan, Frank, Ian, Justin, Ryan, Andrew, Fred, Jack, Matthew, Stephen, Brad, Greg, Jed, Paul, Todd, Brandon, Hank, Jonathan, Peter, Wilbur, Amanda, Courtney, Heather, Melanie, Sara, Amber, Crystal, Katie, Meredith, Shannon, Betsy, Donna, Kristin, Nancy, Stephanie, Ellen, Lauren, Peggy, Colleen, Emily, Megan, Rachel, Wendy, *Bobbie-Sue, Sue-Ellen*}

A_2 = {Alonzo, Jamel, Lerone, Theo, Alphonse, Jerome, Leroy, Torrance, Darnell, Lamar, Lionel, Rashaun, Tyree, Deion, Lamont, Malik, Terrence, Tyrone, Lavon, Marcellus, Nichelle, Shereen, Ebony, Latisha, Shaniqua, Jasmine, Latonya, Tanisha, Tia, Lakisha, Latoya, Yolanda, Malika, Tawanda, Yvette, *Percell, Rasaan, Everol, Terryl, Wardell, Aiesha, Lashelle, Temeka, Tameisha, Teretha, Shanise, Sharise, Tashika, Lashandra, Shavonn*}

Note: Names in italics were removed and **not** replaced due to lack of occurrence in at least one corpus. This highlights the flexibility of the method as the cardinality of the two sets is unequal.

B.4 Target Words

T_1 = {intelligent, bright, smart, clever, educated, knowledgeable, literate, trained, skilled, creative, rich, diligent}

T_2 = {unintelligent, dim, stupid, dense, uneducated, ignorant, illiterate, untrained, unskilled, uncreative, poor, lazy}