Nurse is Closer to Woman than Surgeon? Mitigating Gender-Biased Proximities in Word Embeddings

Vaibhav Kumar[1]* Tenzin Singhay Bhotia[1]* Vaibhav Kumar[1]* Tanmoy Chakraborty[2]

[1]Delhi Technological University, New Delhi, India
[2]IIIT-Delhi, India
[1]{kumar.vaibhav1o1, tenzinbhotia0, vaibhavk992}@gmail.com
[2]tanmoy@iiitd.ac.in

Abstract

Word embeddings are the standard model for semantic and syntactic representations of words. Unfortunately, these models have been shown to exhibit undesirable word associations resulting from gender, racial, and religious biases. Existing post-processing methods for debiasing word embeddings are unable to mitigate gender bias hidden in the spatial arrangement of word vectors. In this paper, we propose **RAN-Debias**, a novel gender debiasing methodology that not only eliminates the bias present in a word vector but also alters the spatial distribution of its neighboring vectors, achieving a bias-free setting while maintaining minimal semantic offset. We also propose a new bias evaluation metric, **Gender-based Illicit Proximity Estimate** (GIPE), which measures the extent of undue proximity in word vectors resulting from the presence of gender-based predilections. Experiments based on a suite of evaluation metrics show that RAN-Debias significantly outperforms the state-of-the-art in reducing proximity bias (GIPE) by at least 42.02%. It also reduces direct bias, adding minimal semantic disturbance, and achieves the best performance in a downstream application task (coreference resolution).

1 Introduction

Word embedding methods (Devlin et al., 2019; Mikolov et al., 2013a; Pennington et al., 2014) have been staggeringly successful in mapping the semantic space of words to a space of real-valued vectors, capturing both semantic and syntactic relationships. However, as recent research has shown, word embeddings also possess a spectrum of biases related to gender (Bolukbasi et al., 2016; Hoyle et al., 2019), race, and religion (Manzini et al., 2019; Otterbacher et al., 2017). Bolukbasi et al. (2016) showed that there is a disparity in the association of professions with gender. For instance, while women are associated more closely with ''receptionist'' and ''nurse'', men are associated more closely with ''doctor'' and ''engineer''. Similarly, a word embedding model trained on data from a popular social media platform generates analogies such as ''Muslim is to terrorist as Christian is to civilian'' (Manzini et al., 2019). Therefore, given the large scale use of word embeddings, it becomes cardinal to remove the manifestation of biases. In this work, we focus on mitigating gender bias from pre-trained word embeddings.

As shown in Table 1, the high degree of similarity between gender-biased words largely results from their individual proclivity towards a particular notion (gender in this case) rather than from empirical utility; we refer to such proximities as ''illicit proximities''. Existing debiasing methods (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019) are primarily concerned with debiasing a word vector by minimising its projection on the gender direction. Although they successfully mitigate direct bias for a word, they tend to ignore the relationship between a gender-neutral word vector and its neighbors, thus failing to remove the gender bias encoded as illicit proximities between words (Gonen and Goldberg, 2019; Williams et al., 2019). For the sake of brevity, we refer to ''gender-based illicit proximities'' as ''illicit proximities'' in the rest of the paper.

*Authors have contributed equally.

Transactions of the Association for Computational Linguistics, vol. 8, pp. 486–503, 2020. https://doi.org/10.1162/tacl_a_00327
Action Editor: Xiaodong He. Submission batch: 1/2020; Revision batch: 5/2020; Published 8/2020.

Word	Neighbors
nurse	mother$_{12}$, woman$_{24}$, filipina$_{31}$
receptionist	housekeeper$_{9}$, hairdresser$_{15}$, prostitute$_{69}$
prostitute	housekeeper$_{19}$, hairdresser$_{41}$, babysitter$_{44}$
schoolteacher	homemaker$_{2}$, housewife$_{4}$, waitress$_{8}$

Table 1: Words and their neighbors extracted using GloVe (Pennington et al., 2014). Subscript indicates the rank of the neighbor.

To account for these problems, we propose a post-processing based debiasing scheme for non-contextual word embeddings, called **RAN-Debias** (**R**epulsion, **A**ttraction, and **N**eutralization based **Debias**ing). RAN-Debias not only minimizes the projection of gender-biased word vectors on the gender direction but also reduces the semantic similarity with neighboring word vectors having illicit proximities. We also propose **KBC** (**K**nowledge **B**ased **C**lassifier), a word classification algorithm for selecting the set of words to be debiased. KBC utilizes a set of existing lexical knowledge bases to maximize classification accuracy. Additionally, we propose a metric, **Gender-based Illicit Proximity Estimate (GIPE)**, which quantifies gender bias in the embedding space resulting from the presence of illicit proximities between word vectors.

We evaluate debiasing efficacy on various evaluation metrics. For the gender relational analogy test on the SemBias dataset (Zhao et al., 2018b), RAN-GloVe (RAN-Debias applied to GloVe word embedding) outperforms the next best baseline GN-GloVe (debiasing method proposed by Zhao et al. [2018b]) by 21.4% in gender-stereotype type. RAN-Debias also outperforms the best baseline by at least 42.02% in terms of GIPE. Furthermore, the performance of RAN-GloVe on word similarity and analogy tasks on a number of benchmark datasets indicates the addition of minimal semantic disturbance. In short, our major contributions[1] can be summarized as follows:

- We provide a knowledge-based method (KBC) for classifying words to be debiased.
- We introduce RAN-Debias, a novel approach to reduce both direct and gender-based proximity biases in word embeddings.
- We propose GIPE, a novel metric to measure the extent of undue proximities in word embeddings.

2 Related Work

2.1 Gender Bias in Word Embedding Models

Caliskan et al. (2017) highlighted that human-like semantic biases are reflected through word embeddings (such as GloVe [Pennington et al., 2014]) of ordinary language. They also introduced the Word Embedding Association Test (WEAT) for measuring bias in word embeddings. The authors showed a strong presence of biases in pre-trained word vectors. In addition to gender, they also identified bias related to race. For instance, European-American names are more associated with pleasant terms as compared to African-American names.

In the following subsections, we discuss existing gender debiasing methods based on their mode of operation. Methods that operate on pre-trained word embeddings are known as *post-processing methods*, and those which aim to retrain word embeddings by either introducing corpus-level changes or modifying the training objective are known as *learning-based methods*.

2.2 Debiasing Methods (Post-processing)

Bolukbasi et al. (2016) extensively studied gender bias in word embeddings and proposed two debiasing strategies—‘‘hard debias’’ and ‘‘soft debias’’. Hard debias algorithm first determines the direction that captures the gender information in the word embedding space using the difference vectors (e.g., $\vec{he} - \vec{she}$). It then transforms each word vector $\vec{w}$ to be debiased such that it becomes perpendicular to the gender direction (neutralization). Further, for a given set of word pairs (equalization set), it modifies each pair such that $\vec{w}$ becomes equidistant to each word in the pair (equalization). On the other hand, the soft debias

[1] The code and data are released at `https://github.com/TimeTraveller-San/RAN-Debias`.

algorithm applies a linear transformation to word vectors, which preserves pairwise inner products among all the word vectors while limiting the projection of gender-neutral words on the gender direction. The authors showed that the former performs better for debiasing than the latter. However, to determine the set of words for debiasing, a support vector machine (SVM) classifier is used, which is trained on a small set of seed words. This makes the accuracy of the approach highly dependent on the generalization of the classifier to all remaining words in the vocabulary.

Kaneko and Bollegala (2019) proposed a post-processing step in which the given vocabulary is split into four classes—non-discriminative female-biased words (e.g., ‘‘bikini’’, ‘‘lipstick’’), non-discriminative male-biased words (e.g., ‘‘beard’’, ‘‘moustache’’), gender-neutral words (e.g., ‘‘meal’’, ‘‘memory’’), and stereotypical words (e.g., ‘‘librarian’’, ‘‘doctor’’). A set of seed words is then used for each of the categories to train an embedding using an encoder in a denoising autoencoder, such that gender-related biases from stereotypical words are removed, while preserving feminine information for non-discriminative female-biased words, masculine information for non-discriminative male-biased words, and neutrality of the gender-neutral words. The use of the correct set of seed words is critical for the approach. Moreover, inappropriate associations between words (such as ‘‘nurse’’ and ‘‘receptionist’’) may persist.

Gonen and Goldberg (2019) showed that current approaches (Bolukbasi et al., 2016; Zhao et al., 2018b), which depend on gender direction for the definition of gender bias and directly target it for the mitigation process, end up hiding the bias rather than reduce it. The relative spatial distribution of word vectors before and after debiasing is similar, and bias-related information can still be recovered.

Ethayarajh et al. (2019) provided theoretical proof for hard debias (Bolukbasi et al., 2016) and discussed the theoretical flaws in WEAT by showing that it systematically overestimates gender bias in word embeddings. The authors presented an alternate gender bias measure, called RIPA (Relational Inner Product Association), that quantifies gender bias using gender direction. Further, they illustrated that vocabulary selection for gender debiasing is as crucial as the debiasing procedure.

Zhou et al. (2019) investigated the presence of gender bias in bilingual word embeddings and languages which have grammatical gender (such as Spanish and French). Further, they defined semantic gender direction and grammatical gender direction used for quantifying and mitigating gender bias. In this paper, we only focus on languages that have non-gendered grammar (e.g., English). Our method can be applied to any such language.

2.3 Debiasing Methods (Learning-based)

Zhao et al. (2018b) developed a word vector training approach, called Gender-Neutral Global Vectors (GN-GloVe) based on the modification of GloVe. They proposed a modified objective function that aims to confine gender-related information to a sub-vector. During the optimization process, the objective function of GloVe is minimized while simultaneously, the square of Euclidean distance between the gender-related sub-vectors is maximized. Further, it is emphasized that the representation of gender-neutral words is perpendicular to the gender direction. Being a retraining approach, this method cannot be used on pre-trained word embeddings.

Lu et al. (2018) proposed a counterfactual data-augmentation (CDA) approach to show that gender bias in language modeling and coreference resolution can be mitigated through balancing the corpus by exchanging gender pairs like ‘‘she’’ and ‘‘he’’ or ‘‘mother’’ and ‘‘father’’. Similarly, Hall Maudslay et al. (2019) proposed a learning-based approach with two enhancements to CDA—a counterfactual data substitution method which makes substitutions with a probability of 0.5 and a method for processing first names based upon bipartite graph matching.

Bordia and Bowman (2019) proposed a gender-bias reduction method for word-level language models. They introduced a regularization term that penalizes the projection of word embeddings on the gender direction. Further, they proposed metrics to measure bias at embedding and corpus level. Their study revealed considerable gender bias in Penn Treebank (Marcus et al., 1993) and WikiText-2 (Merity et al., 2018).

2.4 Word Embeddings Specialization

Mrkšić et al. (2017) defined semantic specialization as the process of refining word vectors to improve the semantic content. Similar to the debiasing procedures, semantic specialization procedures can also be divided into *post-processing* (Ono et al., 2015; Faruqui and Dyer, 2014) and *learning-based* (Rothe and Schütze, 2015; Mrkšić et al., 2016; Nguyen et al., 2016) approaches. The performance of post-processing based approaches is shown to be better than learning-based approaches (Mrkšić et al., 2017).

Similar to the ''repulsion'' and ''attraction'' terminologies used in RAN-Debias, Mrkšić et al. (2017) defined ATTRACT-REPEL algorithm, a post-processing semantic specialization process which uses antonymy and synonymy constraints drawn from lexical resources. Although it is superficially similar to RAN-Debias, there are a number of differences between the two approaches. Firstly, the ATTRACT-REPEL algorithm operates over mini-batches of synonym and antonym pairs, while RAN-Debias operates on a set containing gender-neutral and gender-biased words. Secondly, the ''attract'' and ''repel'' terms carry different meanings with respect to the algorithms. In ATTRACT-REPEL, for each of the pairs in the mini-batches of synonyms and antonyms, negative examples are chosen. The algorithm then forces synonymous pairs to be closer to each other (attract) than from their negative examples and antonymous pairs further away from each other (repel) than from their negative examples. On the other hand, for a given word vector, RAN-Debias forces it away from its neighboring word vectors (repel) which have a high indirect bias while simultaneously forcing the post-processed word vector and the original word vector together (attract) to preserve its semantic properties.

3 Proposed Approach

Given a set of pre-trained word vectors $\{\vec{w}_i\}_{i=1}^{|V|}$ over a vocabulary set V, we aim to create a transformation $\{\vec{w}_i\}_{i=1}^{|V|} \rightarrow \{\vec{w'}_i\}_{i=1}^{|V|}$ such that the stereotypical gender information present in the resulting embedding set are minimized with minimal semantic offset. We first define the categories into which each word $w \in V$ is classified in a mutually exclusive manner. Table 2 summarizes important notations used throughout the paper.

Notation	Denotation
$\vec{w}$	Vector corresponding to a word w
$\vec{w_d}$	Debiased version of $\vec{w}$
V	Vocabulary set
V_p	The set of words which are preserved during the debiasing procedure
V_d	The set of words which are subjected to the debiasing procedure
D	Set of dictionaries
d_i	A particular dictionary from the set D
$\vec{g}$	Gender direction
$D_b(\vec{w})$	Direct bias of a word w.
$\beta(\vec{w_1}, \vec{w_2})$	Indirect bias between a pair of words w_1 and w_2.
$\eta(\vec{w})$	Gender-based proximity bias of a word w
N_w	Set of neighboring words of a word w
$F_r(\vec{w_d})$	Repulsion objective function
$F_a(\vec{w_d})$	Attraction objective function
$F_n(\vec{w_d})$	Neutralization objective function
$F(\vec{w_d})$	Multi-objective optimization function
KBC	Knowledge Based Classifier
BBN	Bias Based Network
$GIPE$	Gender-based Illicit Proximity Estimate

Table 2: Important notations and denotations.

- **Preserve set** (V_p): This set consists of words for which gender carries semantic importance; such as names, gendered pronouns and words like ''beard'' and ''bikini'' that have a meaning closely associated with gender. In addition, words that are non-alphabetic are also included as debiasing them will be of no practical utility.

- **Debias set** (V_d): This set consists of all the words in the vocabulary that are not present in V_p. These words are expected to be gender-neutral in nature and hence subjected to debiasing procedure. Note that V_d not only consists of gender-stereotypical words (''nurse'', ''warrior'', ''receptionist'', etc.), but also gender-neutral words (''sky'', ''table'', ''keyboard'', etc.).

3.1 Word Classification Methodology

Prior to the explanation of our method, we present the limitations of previous approaches for word classification. Bolukbasi et al. (2016) trained a linear SVM using a set of gender-specific seed words, which is then generalized on the whole embedding set to identify other gender-specific

Method	Prec	Rec	F1	AUC-ROC	Acc
SVM	**97.20**	59.37	73.72	83.98	78.83
RIPA	60.60	53.40	56.79	59.51	59.35
KBC	89.65	**81.25**	**85.24**	**86.25**	**85.93**

Table 3: Comparison between our proposed method (KBC), RIPA- (Ethayarajh et al., 2019), and SVM- (Bolukbasi et al., 2016) based word classification methods via precision (Prec), recall (Rec), F1-score (F1), AUC-ROC, and accuracy (Acc).

words. However, such methods rely on training a supervised classifier on word vectors, which are themselves gender-biased. Because such classifiers are trained on biased data, they catch onto the underlying gender-bias cues and often misclassify words. For instance, the SVM classifier trained by Bolukbasi et al. (2016) misclassifies the word ‘‘blondes’’ as gender-specific, among others. Further, we empirically show the inability of a supervised classifier (SVM) to generalize over the whole embedding using various metrics in Table 3.

Taking into consideration this limitation, we propose the Knowledge Based Classifier (KBC) that relies on knowledge bases instead of word embeddings, thereby circumventing the addition of bias in the classification procedure. Moreover, unlike RIPA (Ethayarajh et al., 2019), our approach does not rely on creating a biased direction that may be difficult to determine. Essentially, KBC relies on the following assumption.

Assumption 1 *If there exists a dictionary* d *such that it stores a definition* $d[w]$ *corresponding to a word* w*, then* w *can be defined as gender-specific or not based on the existence or absence of a gender-specific reference* $s \in seed$ *in the definition* $d[w]$*, where the set seed consists of gender-specific references such as* {*‘‘man’’, ‘‘woman’’, ‘‘boy’’, ‘‘girl’’*}.

Algorithm 1: Knowledge Based Classifier (KBC)

```
Input  : $V$: vocabulary set, $isnonaphabetic(w)$:
         checks for non-alphabetic words
         $seed$: set of gender-specific words
         $stw$: set of stop words
         $names$: set of gender-specific names
         $D$: set of dictionaries, where for a
         dictionary $d_i \in D$, $d_i[w]$ represents the
         definition of a word $w$.
Output: $V_p$: set of words that will be preserved,
        $V_d$: set of words that will be debiased
$V_p = \{\}, V_d = \{\}$
for $w \in V$ do
    if $w \in$ stw or isnonalphabetic(w) then
        $V_p \leftarrow V_p \cup \{w\}$
    else if $w \in names \cup seed$ then
        $V_p \leftarrow V_p \cup \{w\}$
    else if $|\{d_i : d_i \in D \;\&$
      $w \in d_i \;\& \;\exists s : s \in seed \cap d_i[w]\}| > |D|/2$
      then
        $V_p \leftarrow V_p \cup \{w\}$
$V_d \leftarrow V_d \cup \{w : w \in V \setminus V_p\}$
return $V_p, V_d$
```

Algorithm 1 formally explains KBC. We denote each *if condition* as a stage and explain it below:

- **Stage 1:** This stage classifies all stop words and non-alphabetic words as V_p. Debiasing such words serve no practical utility; hence we preserve them.

- **Stage 2:** This stage classifies all words that belong to either $names$ set or $seed$ set as V_p. Set $names$ is collected from open source knowledge base.[2] Set $seed$ consists of gender-specific reference terms. We preserve names, as they hold important gender information (Pilcher, 2017).

- **Stage 3:** This stage uses a collection of dictionaries to determine whether a word is gender-specific using Assumption 1. To counter the effect of biased definitions arising from any particular dictionary, we make a decision based upon the consensus of all dictionaries. A word is classified as gender-specific and added to V_p if and only if more than half of the dictionaries classify it as gender-specific. In our experiments, we employ WordNet (Miller, 1995) and

[2] `https://github.com/ganoninc/fb-gender-json`.

the Oxford dictionary. As pointed out by Bolukbasi et al. (2016), WordNet consists of few definitions that are gender-biased such as the definition of ''vest''; therefore, by utilizing our approach, we counter such cases as the final decision is based upon consensus.

The remaining words that are not preserved by KBC are categorized into V_d. It is the set of words that are debiased by RAN-Debias later.

3.2 Types of Gender Bias

First, we briefly explain two types of gender bias as defined by Bolukbasi et al. (2016) and then introduce a new type of gender bias resulting from illicit proximities in word embedding space.

- **Direct Bias** (D_b): For a word w, the direct bias is defined by

$$D_b(\vec{w}, \vec{g}) = |cos(\vec{w}, \vec{g})|^c$$

where, $\vec{g}$ is the gender direction measured by taking the first principal component from the principal component analysis of ten gender pair difference vectors, such as $(\vec{he} - \vec{she})$ as mentioned in (Bolukbasi et al., 2016), and c represents the strictness of measuring bias.

- **Indirect Bias** (β): The indirect bias between a given pair of words w and v is defined by

$$\beta(\vec{w}, \vec{v}) = \frac{(\vec{w}.\vec{v} - cos(\vec{w}_\perp, \vec{v}_\perp))}{\vec{w}.\vec{v}}$$

Here, $\vec{w}$ and $\vec{v}$ are normalized. $\vec{w}_\perp$ is orthogonal to the gender direction $\vec{g}$: $\vec{w}_\perp = \vec{w} - \vec{w}_g$, and $\vec{w}_g$ is the contribution from gender: $\vec{w}_g = (\vec{w}.\vec{g})\vec{g}$. Indirect bias measures the change in the inner product of two word vectors as a proportion of the earlier inner product after projecting out the gender direction from both the vectors. A higher indirect bias between two words indicates a strong association due to gender.

- **Gender-based Proximity Bias** (η): Gonen and Goldberg (2019) observed that the existing debiasing methods are unable to completely debias word embeddings because the relative spatial distribution of word embeddings after the debiasing process still encapsulates bias-related information. Therefore, we propose gender-based proximity bias that aims to capture the illicit proximities arising between a word and its closest k neighbors due to gender-based constructs. For a given word $w_i \in V_d$, the gender-based proximity bias η_{w_i} is defined as:

$$\eta_{w_i} = \frac{|N^b_{w_i}|}{|N_{w_i}|} \tag{1}$$

where

$$N_{w_i} = \underset{V':|V'|=k}{\text{argmax}} \left(\cos(\vec{w_i}, \vec{w_k}) : w_k \in V' \subseteq V\right),$$
$$N^b_{w_i} = \{w_i : \beta(\vec{w_i}, \vec{w_k}) > \theta_s,\ w_k \in N_{w_i}\}, \text{and } \theta_s$$

is a threshold for indirect bias.

The intuition behind this is as follows. The set N_{w_i} consists of the top k neighbors of w_i calculated by finding the word vectors having the maximum cosine similarity with w_i. Further, $N^b_{w_i} \subseteq N_{w_i}$ is the set of neighbors having indirect bias β greater than a threshold θ_s, which is a hyperparameter that controls neighbor deselection on the basis of indirect bias. The lower is the value of θ_s, the higher is the cardinality of set $N^b_{w_i}$. A high value of $|N^b_{w_i}|$ compared to $|N_{w_i}|$ indicates that the neighborhood of the word is gender-biased.

3.3 Proposed Method–RAN-Debias

We propose a multi-objective optimization based solution to mitigate both direct[3] and gender-based proximity bias while adding minimal impact to the semantic and analogical properties of the word embedding. For each word $w \in V_d$ and its vector $\vec{w} \in \mathbb{R}^h$, where h is the embedding dimension, we find its debiased counterpart $\vec{w_d} \in \mathbb{R}^h$ by solving the following multi-objective optimization problem:

$$\underset{\vec{w_d}}{\text{argmin}} \left(F_r(\vec{w_d}), F_a(\vec{w_d}), F_n(\vec{w_d})\right) \tag{2}$$

We solve this by formulating a single objective $F(\vec{w_d})$ and scalarizing the set of objectives using the weighted sum method as follows:

$$F(\vec{w_d}) = \lambda_1.F_r(\vec{w_d}) + \lambda_2.F_a(\vec{w_d}) + \lambda_3.F_n(\vec{w_d})$$
$$\text{such that } \lambda_i \in [0,1] \text{ and } \sum_i \lambda_i = 1 \tag{3}$$

$F(\vec{w_d})$ is minimized using the Adam (Kingma and Ba, 2015) optimized gradient descent to obtain the optimal debiased embedding $\vec{w_d}$.

[3] Though not done explicitly, reducing direct bias also reduces indirect bias as stated by Bolukbasi et al. (2016).

As shown in the subsequent sections, the range of objective functions F_r, F_a, F_n (defined later) is $[0, 1]$; thus we use the weights λ_i for determining the relative importance of one objective function over another.

3.3.1 Repulsion

For any word $w \in V_d$, we aim to minimize the gender bias based illicit associations. Therefore, our objective function aims to ‘‘repel’’ $\vec{w}_d$ from the neighboring word vectors which have a high value of indirect bias (β) with it. Consequently, we name it ‘‘repulsion’’ (F_r) and primarily define the repulsion set S_r to be used in F_r as follows.

Definition 1 *For a given word w, the repulsion set S_r is defined as $S_r = \{n_i : n_i \in N_w$ and $\beta(\vec{w}, \vec{n_i}) > \theta_r\}$, where N_w is the set of top 100 neighbors obtained from the original word vector $\vec{w}$.*

Because we aim to reduce the unwanted semantic similarity between $\vec{w}_d$ and the set of vectors S_r, we define the objective function F_r as follows.

$$F_r(\vec{w}_d) = \left(\sum_{n_i \epsilon S_r} \left| cos(\vec{w}_d, \vec{n_i}) \right| \right) \Big/ \left| S_r \right|,$$
$$F_r(\vec{w}_d) \in [0, 1]$$

For our experiments, we find that $\theta_r = 0.05$ is the appropriate threshold to repel majority of gender-biased neighbors.

3.3.2 Attraction

For any word $w \in V_d$, we aim to minimize the loss of semantic and analogical properties for its debiased counterpart $\vec{w}_d$. Therefore, our objective function aims to attract $\vec{w}_d$ towards $\vec{w}$ in the word embedding space. Consequently, we name it ‘‘attraction’’ (F_a) and define it as follows:

$$\begin{aligned} F_a(\vec{w}_d) &= |\cos(\vec{w}_d, \vec{w}) - \cos(\vec{w}, \vec{w})|/2 \\ &= |\cos(\vec{w}_d, \vec{w}) - 1|/2, F_a(\vec{w}_d) \in [0, 1] \end{aligned}$$

3.3.3 Neutralization

For any word $w \in V_d$, we aim to minimize its bias towards any particular gender. Therefore, the objective function F_n represents the absolute value of dot product of word vector $\vec{w}_d$ with the gender direction $\vec{g}$ (as defined by Bolukbasi et al., 2016). Consequently, we name it ‘‘neutralization’’ (F_n) and define it as follows:

$$F_n(\vec{w}_d) = |cos(\vec{w}_d, \vec{g})|, F_n \in [0, 1]$$

3.3.4 Time Complexity of RAN-Debias

Computationally, there are two major components of RAN-Debias:

1. Calculate neighbors for each word $w \in V_d$ and store them in a hash table. This has a time complexity of $O(n^2)$ where $n = |V_d|$.
2. Debias each word using gradient descent, whose time complexity is $O(n)$.

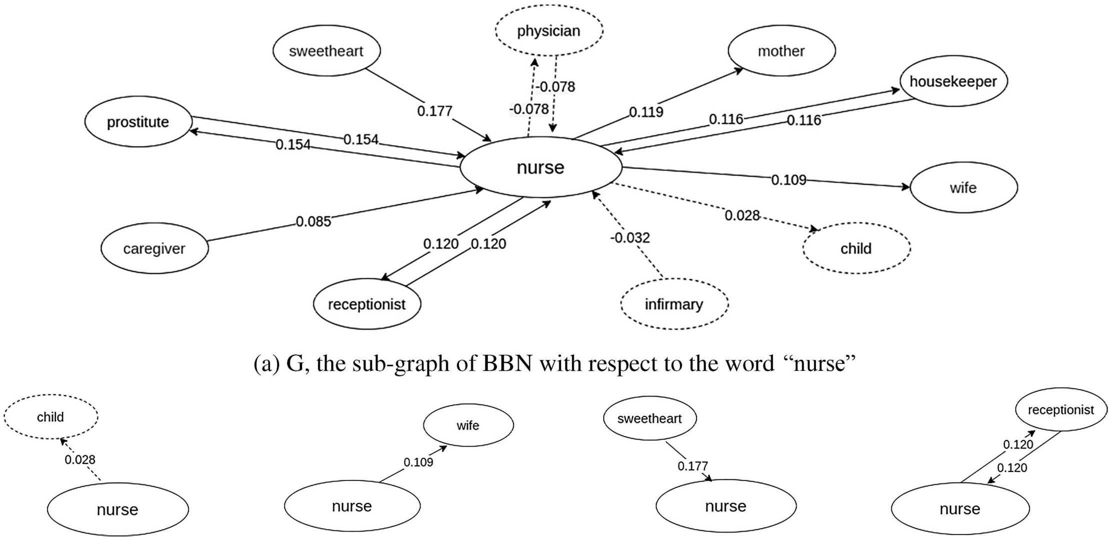
The overall complexity of RAN-Debias is $O(n^2)$, that is, quadratic with respect to the cardinality of debias set V_d.

3.4 Gender-based Illicit Proximity Estimate (GIPE)

In Section 3.2, we defined the gender proximity bias (η). In this section, we extend it to the embedding level for generating a global estimate. Intuitively, an estimate can be generated by simply taking the mean of η_w, $\forall w \in V_d$. However, this computation assigns equal importance to all η_w values, which is an oversimplification. A word w may itself be in the proximity of another word $w' \in V_d$ through gender-biased associations, thereby increasing $\eta_{w'}$. Such cases in which w increases $\eta_{w'}$ for other words should also be taken into account. Therefore, we use a weighted average of η_w, $\forall w \in V$ for determining a global estimate. We first define a weighted directed network, called **Bias Based Network (BBN)**. The use of a graph data structure makes it easier to understand the intuition behind GIPE.

Definition 2 *Given a set of non gender-specific words W, bias based network is a directed graph $G = (V, E)$, where nodes represent word vectors and weights of directed edges represent the indirect bias (β) between them. The vertex set V and edge set E are obtained according to Algorithm 2.*

For each word w_i in W, we find N, the set of top n word vectors having the highest cosine similarity with $\vec{w_i}$ (we keep n to be 100 to reduce computational overhead without compromising on quality). For each pair $(\vec{w_i}, \vec{w_k})$, where $w_k \in N$, a directed edge is assigned from w_i to w_k with the edge weight being $\beta(\vec{w_i}, \vec{w_k})$. In case the given



(a) G, the sub-graph of BBN with respect to the word "nurse"

(b) Edge with $\beta < \theta_s$ (c) Outgoing edge (d) Incoming edge (e) Dual edges

Figure 1: (a): A sub-graph of BBN formed by Algorithm 2 for GloVe (Pennington et al., 2014) trained on 2017-January dump of Wikipedia; we discuss the structure of the graph with respect to the word ''nurse''. We illustrate four possible scenarios with respect to their effect on GIPE, with θ_s = 0.05: (b) An edge with $\beta < \theta_s$ may not contribute to γ_i or η_{w_i}; (c) An outgoing edge may contribute to η_{w_i} only; (d) An incoming edge may contribute to γ_i only; (e) Incoming and outgoing edges may contribute to γ_i and η_{w_i} respectively. Every node pair association can be categorized as one of the aforementioned four cases.

Algorithm 2: Compute BBN for the given set of word vectors

```
Input  : $\xi$: word embedding set,
         $W$: set of non gender-specific words,
         $n$: number of neighbors
Output: $G$: bias based network
$V = [\,], E = [\,]$
for $x_i \in W$ do
    $N = \underset{\xi' : |\xi'| = n}{\text{argmax}} \left(\cos(\vec{x_i}, \vec{x_k}) : x_k \in \xi' \subseteq \xi\right)$
    $V.insert(x_i)$
    for $x_k \in N$ do
        $E.insert\left(x_i, x_k, \beta\left(\vec{x_i}, \vec{x_k}\right)\right)$
        $V.insert\left(x_k\right)$
G = (V, E)
return $G$
```

embedding is a debiased version, we use the non-debiased version of the embedding for computing $\beta(\vec{w_i}, \vec{w_k})$. Figure 1 portrays a sub-graph in BBN. By representing the set of non gender-specific words as a weighted directed graph we can use the number of outgoing and incoming edges for a node (word w_i) for determining η_{w_i} and its weight respectively, thereby leading to the formalization of GIPE as follows.

Definition 3 *For a BBN G, the Gender-based Illicit Proximity Estimate of G, indicated by $GIPE(G)$ is defined as:*

$$GIPE(G) = \frac{\sum_{i=1}^{|V|} \gamma_i \eta_{w_i}}{\sum_{i=1}^{|V|} \gamma_i}$$

where, for a word w_i, η_{w_i} is the gender-based proximity bias as defined earlier, ϵ is a (small) positive constant, and γ_i is the weight, defined as:

$$\gamma_i = 1 + \frac{|\{v_i : (v_i, w_i) \in E, \beta(\vec{v_i}, \vec{w_i}) > \theta_s\}|}{\epsilon + |\{v_i : (v_i, w_i) \in E\}|} \tag{4}$$

The intuition behind the metric is as follows. For a bias based network G, $GIPE(G)$ is the weighted average of gender-based proximity bias (η_{w_i}) for all nodes $w_i \in W$, where the weight of a node is γ_i, which signifies the importance of the node in contributing towards the gender-based proximity bias of other word vectors. γ_i takes into account the number of incoming edges having β higher than a threshold θ_s. Therefore, we take into account how the neighborhood of a node contributes towards illicit proximities (having high β values for outgoing edges) as well as how a node itself contributes towards illicit proximities of other nodes (having high β values for incoming edges). For illustration, we analyze a sub-graph in Figure 1. By incorporating γ_i, we take into account both dual (Figure 1e) and incoming (Figure 1d) edges, which would not have been the case otherwise. In GloVe (2017-January dump of Wikipedia), the word ''sweetheart'' has ''nurse'' in the set of its top 100 neighbors and $\beta > \theta_s$; however, ''nurse'' does not have ''sweetheart'' in the set of its top 100 neighbors. Hence, while ''nurse'' contributes towards gender-based proximity bias of the word ''sweetheart'', vice versa is not true. Similarly, if dual-edge exists, then both γ_i and η_{w_i} are taken into account. Therefore, GIPE considers all possible cases of edges in BBN, making it a holistic metric.

4 Experiment Results

We conduct the following performance evaluation tests:

- We compare KBC with SVM-based (Bolukbasi et al., 2016) and RIPA-based (Ethayarajh et al., 2019) methods for word classification.
- We evaluate the capacity of RAN-Debias on GloVe (*aka* RAN-GloVe) for the gender relational analogy dataset–SemBias (Zhao et al., 2018b).
- We demonstrate the ability of RAN-GloVe to mitigate gender proximity bias by computing and contrasting the GIPE value.
- We evaluate RAN-GloVe on several benchmark datasets for similarity and analogy tasks, showing that RAN-GloVe introduces minimal semantic offset to ensure quality of the word embeddings.
- We demonstrate that RAN-GloVe successfully mitigates gender bias in a downstream application - coreference resolution.

Although we report and analyze the performance of RAN-GloVe in our experiments, we also applied RAN-Debias to other popular non-contextual and monolingual word embedding, Word2vec (Mikolov et al., 2013a) to create RAN-Word2vec. As expected, we observed similar results (hence not reported for the sake of brevity), emphasizing the generality of RAN-Debias. Note that the percentages mentioned in the rest of the section are relative unless stated otherwise.

4.1 Training Data and Weights

We use GloVe (Pennington et al., 2014) trained on the 2017-January dump of Wikipedia, consisting of 322,636 unique word vectors of 300 dimensions. We apply KBC on the vocabulary set V obtaining V_p and V_d of size 47,912 and 274,724 respectively. Further, judging upon the basis of performance evaluation tests as discussed above, we experimentally select the weights in Equation 3 as $\lambda_1 = 1/8, \lambda_2 = 6/8,$ and $\lambda_3 = 1/8.$

4.2 Baselines for Comparisons

We compare RAN-GloVe against the following word embedding models, each of which is trained on the 2017-January dump of Wikipedia.

- **GloVe**: A pre-trained word embedding model as mentioned earlier. This baseline represents the non-debiased version of word embeddings.
- **Hard-GloVe**: Hard-Debias GloVe; we use the debiasing method[4] proposed by Bolukbasi et al., 2016 on GloVe.
- **GN-GloVe**: Gender-neutral GloVe; we use the original[5] debiased version of GloVe released by Zhao et al. (2018b).
- **GP-GloVe**: Gender-preserving GloVe; we use the original[6] debiased version of GloVe released by Kaneko and Bollegala (2019).

4.3 Word Classification

We compare KBC with RIPA-based (unsupervised) (Ethayarajh et al., 2019) and SVM-based (supervised) (Bolukbasi et al., 2016) approaches for word classification. We create a balanced labeled test set consisting of a total of 704 words, with 352 words for each category—gender-specific and non gender-specific. For the non gender-specific category, we select all the 87 neutral and biased words from the SemBias dataset (Zhao et al., 2018b). Further, we select all 320, 40 and 60 gender-biased occupation words released by Bolukbasi et al. (2016); Zhao et al. (2018a) and Rudinger et al. (2018), respectively. After combining and removing duplicate words, we obtain

[4] https://github.com/tolga-b/debiaswe.

[5] https://github.com/uclanlp/gn_GloVe.

[6] https://github.com/kanekomasahiro/gp_debias.

Dataset	Embedding	Definition ↑	Stereotype ↓	None ↓
SemBias	GloVe	80.2	10.9	8.9
	Hard-GloVe	84.1	6.4	9.5
	GN-GloVe	**97.7**	1.4	**0.9**
	GP-GloVe	84.3	8.0	7.7
	RAN-GloVe	92.8	**1.1**	6.1

Table 4: Comparison for the gender relational analogy test on the SemBias dataset. ↑ (↓) indicates that higher (lower) value is better.

352 non gender-specific words. For the gender-specific category, we use a list of 222 male and 222 female words provided by Zhao et al. (2018b). We use stratified sampling to under-sample 444 words into 352 words for balancing the classes. The purpose of creating this diversely sourced dataset is to provide a robust ground-truth for evaluating the efficacy of different word classification algorithms.

Table 3 shows precision, recall, F1-score, AUC-ROC, and accuracy by considering gender-specific words as the positive class and non gender-specific words as the negative class. Thus, for KBC, we consider the output set V_p as the positive and V_d as the negative class.

The SVM-based approach achieves high precision but at the cost of a low recall. Although the majority of the words classified as gender-specific are correct, it achieves this due to the limited coverage of the rest of gender-specific words, resulting in them being classified as non gender-specific, thereby reducing the recall drastically.

The RIPA approach performs fairly with respect to precision and recall. Unlike SVM, RIPA is not biased towards a particular class and results in rather fair performance for both the classes. Almost similar to SVM, KBC also correctly classifies most of the gender-specific words but in an exhaustive manner, thereby leading to much fewer misclassification of gender-specific words as non gender-specific. As a result, KBC achieves sufficiently high recall.

Overall, KBC outperforms the best baseline by an improvement of 2.7% in AUC-ROC, 15.6% in F1-score, and 9.0% in accuracy. Additionally, because KBC entirely depends on knowledge bases, the absence of a particular word in them may result in misclassification. This could be the reason behind the lower precision of KBC as compared to SVM-based classification and can be improved upon by incorporating more extensive knowledge bases.

4.4 Gender Relational Analogy

To evaluate the extent of gender bias in RAN-GloVe, we perform gender relational analogy test on the SemBias (Zhao et al., 2018b) dataset. Each instance of SemBias contains four types of word pairs: a gender-definition word pair (**Definition**; ‘‘headmaster-headmistress’’), a gender-stereotype word pair (**Stereotype**; ‘‘manager-secretary’’) and two other word pairs which have similar meanings but no gender-based relation (**None**; ‘‘treble - bass’’). There are a total of 440 instances in the semBias dataset, created by the cartesian product of 20 gender-stereotype word pairs and 22 gender-definition word pairs. From each instance, we select a word pair (a, b) from the four word pairs such that using the word embeddings under evaluation, cosine similarity of the word vectors $(\vec{he} - \vec{she})$ and $(\vec{a} - \vec{b})$ would be maximum. Table 4 shows an embedding-wise comparison on the SemBias dataset. The accuracy is measured in terms of the percentage of times each type of word pair is selected as the top for various instances. RAN-GloVe outperforms all other post-processing debiasing methods by achieving at least 9.96% and 82.8% better accuracy in gender-definition and gender-stereotype, respectively. We attribute this performance to be an effect of superior vocabulary selection by KBC and the neutralization objective of RAN-Debias. KBC classifies the words to be debiased or preserved with high accuracy, while the neutralization objective function of RAN-Debias directly minimizes the preference of a biased word between ‘‘he’’ and ‘‘she’’; reducing the gender cues that give rise to unwanted gender-biased analogies (Table 10). Therefore, although RAN-GloVe achieves lower accuracy for gender-definition type as compared to (learning-based)

Input	Embedding	GIPE		
		$\theta_s = 0.03$	$\theta_s = 0.05$	$\theta_s = 0.07$
V_d	GloVe	0.115	0.038	0.015
	Hard-GloVe	0.069	0.015	0.004
	GN-GloVe	0.142	0.052	0.022
	GP-GloVe	0.145	0.048	0.018
	RAN-GloVe	**0.040**	**0.006**	**0.002**
H_d	GloVe	0.129	0.051	0.024
	Hard-GloVe	0.075	0.020	**0.007**
	GN-GloVe	0.155	0.065	0.031
	GP-GloVe	0.157	0.061	0.027
	RAN-GloVe	**0.056**	**0.018**	0.011

Table 5: GIPE (range: 0–1) for different values of θ_s (lower value is better).

GN-GloVe, it outperforms the next best baseline in **Stereotype** by at least 21.4%.

4.5 Gender-based Illicit Proximity Estimate

GIPE analyzes the extent of undue gender bias based proximity between word vectors. An embedding-wise comparison for various values of θ_s is presented in Table 5. For a fair comparison, we compute GIPE for a BBN created upon our debias set V_d as well as for H_d, the set of words debiased by Bolukbasi et al. (2016).

Here, θ_s represents the threshold as defined earlier in Equation 4. As it may be inferred from Equations 1 and 4, upon increasing the value of θ_s, for a word w_i, the value of both η_{w_i} and γ_i decreases, as a lesser number of words qualifies the threshold for selection in each case. Therefore, as evident from Table 5, the value of GIPE decreases with the increase of θ_s.

For the input set V_d, RAN-GloVe outperforms the next best baseline (Hard-GloVe) by at least 42.02%. We attribute this to the inclusion of the repulsion objective function F_r in Equation 2, which reduces the unwanted gender-biased associations between words and their neighbors. For the input set H_d, RAN-GloVe performs better than other baselines for all values of θ_s except for $\theta_s = 0.07$ where it closely follows Hard-GloVe.

Additionally, H_d consists of many misclassified gender-specific words, as observed from the low recall performance at the word classification test in Section 4.3. Therefore, the values of GIPE corresponding to every value of θ_s for the input H_d is higher as compared to the values for V_d.

Although there is a significant reduction in GIPE value for RAN-GloVe as compared to other word embedding models, word pairs with noticeable β values still exist (as indicated by non-zero GIPE values), which is due to the tradeoff between semantic offset and bias reduction. As a result, GIPE for RAN-GloVe is not perfectly zero but close to it.

4.6 Analogy Test

The task of analogy test is to answer the following question: ‘‘p is to q as r is to ?’’. Mathematically, it aims at finding a word vector $\vec{w}_{\mathrm{s}}$ which has the maximum cosine similarity with $(\vec{w}_{\mathrm{q}} - \vec{w}_{\mathrm{p}} + \vec{w}_{\mathrm{r}})$. However, Schluter (2018) highlights some critical issues with word analogy tests. For instance, there is a mismatch between the distributional hypothesis used for generating word vectors and the word analogy hypothesis. Nevertheless, following the practice of using word analogy test to ascertain the semantic prowess of word vectors, we evaluate RAN-GloVe to provide a fair comparison with other baselines.

We use Google (Mikolov et al., 2013a) (semantic [Sem] and syntactic [Syn] analogies, containing a total 19,556 questions) and MSR (Mikolov et al., 2013b) (containing a total 7,999 syntactic questions) datasets for evaluating the performance of word embeddings. We use 3CosMul (Levy and Goldberg, 2014) for finding $\vec{w}_{\mathrm{s}}$.

Table 6(a) shows that RAN-GloVe outperforms other baselines on the Google (Sem and Syn) dataset while closely following on the MSR dataset. The improvement in performance can be attributed to the removal of unwanted neighbors of a word vector (having gender bias based proximity), while enriching the neighborhood with those having empirical utility, leading to a better performance in analogy tests.

Embedding	(a) Analogy			(b) Semantic					
	Google-Sem	Google-Syn	MSR	RG	MTurk	RW	MEN	SimLex999	AP
GloVe	79.02	52.26	51.49	75.29	64.27	31.63	72.19	34.86	60.70
Hard-GloVe	80.26	62.76	**51.59**	**76.50**	64.26	31.45	72.19	35.17	59.95
GN-GloVe	76.13	51.00	49.29	74.11	**66.36**	**36.20**	**74.49**	**37.12**	61.19
GP-GloVe	79.15	51.55	48.88	75.30	63.46	27.64	69.78	34.02	57.71
RAN-GloVe	**80.29**	**62.89**	50.98	76.22	64.09	31.33	72.09	34.36	**61.69**

Table 6: Comparison of various embedding methods for (a) analogy tests (performance is measured in accuracy) and (b) word semantic similarity tests (performance is measured in terms of Spearman rank correlation).

4.7 Word Semantic Similarity Test

A word semantic similarity task is a measure of how closely a word embedding model captures the similarity between two words as compared to human-annotated ratings. For a word pair, we compute the cosine similarity between the word embeddings and its Spearman correlation with the human ratings. The word pairs are selected from the following benchmark datasets: RG (Rubenstein and Goodenough, 1965), MTurk (Radinsky et al., 2011), RW (Luong et al., 2013), MEN (Bruni et al., 2014), SimLex999 (Hill et al., 2015), and AP (Almuhareb and Poesio, 2005). The results for these tests are obtained from the word embedding benchmark package (Jastrzebski et al., 2017).[7] Note that it is not our primary aim to achieve a state-of-the-art result in this test. It is only considered to evaluate semantic loss. Table 6(b) shows that RAN-GloVe performs better or follows closely to the best baseline. This shows that RAN-Debias introduces minimal semantic disturbance.

4.8 Coreference Resolution

Finally, we evaluate the performance of RAN-GloVe on a downstream application task—coreference resolution. The aim of coreference resolution is to identify all expressions which refer to the same entity in a given text. We evaluate the embedding models on the OntoNotes 5.0 (Weischedel et al., 2012) and the WinoBias (Zhao et al., 2018a) benchmark datasets. WinoBias comprises sentences constrained by two prototypical templates (Type 1 and Type 2), where each template is further divided into two subsets (PRO and ANTI). Such a construction facilitates in revealing the extent of gender bias present in coreference resolution models. Although both templates are designed to assess the efficacy of coreference resolution models, Type 1 is exceedingly challenging as compared to Type 2 as it has no syntactic cues for disambiguation. Each template consists of two subsets for evaluation—pro-stereotype (PRO) and anti-stereotype (ANTI). PRO consists of sentences in which the gendered pronouns refer to occupations biased towards the same gender. For instance, consider the sentence ‘‘The doctor called the nurse because he wanted a vaccine.’’ Stereotypically, ‘‘doctor’’ is considered to be a male-dominated profession, and the gender of pronoun referencing it (‘‘he’’) is also male. Therefore, sentences in PRO are consistent with societal stereotypes. ANTI consists of the same sentences as PRO, but the gender of the pronoun is changed. Considering the same example but by replacing ‘‘he’’ with ‘‘she’’, we get: ‘‘The doctor called the nurse because she wanted a vaccine.’’ In this case, the gender of pronoun (‘‘she’’) which refers to ‘‘doctor’’ is female. Therefore, sentences in ANTI are not consistent with societal stereotypes. Due to such construction, gender bias in the word embeddings used for training the coreference model would naturally perform better in PRO than ANTI and lead to a higher absolute difference (*Diff*) between them. While a lesser gender bias in the model would attain a smaller *Diff*, the ideal case produces an absolute difference of zero.

Following the coreference resolution testing methodology used by Zhao et al. (2018b), we train the coreference resolution model proposed by Lee et al. (2017) on the OntoNotes train dataset for different embeddings. Table 7 shows F1-score on OntoNotes 5.0 test set, WinoBias PRO and ANTI test set for Type 1 template, along

[7] `https://github.com/kudkudak/word-embeddings-benchmarks`.

Embedding	OntoNotes	PRO	ANTI	*Diff*
GloVe	**66.5**	**76.2**	46.0	30.2
Hard-GloVe	66.2	70.6	54.9	15.7
GN-GloVe	66.2	72.4	51.9	20.5
GP-GloVe	66.2	70.9	52.1	18.8
RAN-GloVe	66.2	61.4	**61.8**	**0.4**

Table 7: F1-Score (in %) in the task of coreference resolution. *Diff* denotes the absolute difference between F1-score on PRO and ANTI datasets.

Input	Embedding	GIPE		
		$\theta_s = 0.03$	$\theta_s = 0.05$	$\theta_s = 0.07$
V_d	AN-GloVe	0.069	0.015	0.004
	RA-GloVe	0.060	0.014	0.007
	RAN-GloVe	**0.040**	**0.006**	**0.002**

Table 8: Ablation study—GIPE for AN-GloVe and RA-GloVe.

with the absolute difference (*Diff*) of F1-scores on PRO and ANTI datasets for different word embeddings. The results for GloVe, Hard-GloVe, and GN-GloVe are obtained from Zhao et al. (2018b).

Table 7 shows that RAN-GloVe achieves the smallest absolute difference between scores on PRO and ANTI subsets of WinoBias, significantly outperforming other embedding models and achieving 97.4% better *Diff* (see Table 7 for the definition of *Diff*) than the next best baseline (Hard-GloVe) and 98.7% better than the original GloVe. This lower *Diff* is achieved by an improved accuracy in ANTI and a reduced accuracy in PRO. We hypothesise that the high performance of non-debiased GloVe in PRO is due to the unwanted gender cues rather than the desired coreference resolving ability of the model. Further, the performance reduction in PRO for the other debiased versions of GloVe also corroborates this hypothesis. Despite debiasing GloVe, a considerable amount of gender cues remain in the baseline models as quantified by a lower, yet significant *Diff*. In contrast, RAN-GloVe is able to remove gender cues dramatically, thereby achieving an almost ideal *Diff*. Additionally, the performance of RAN-GloVe on the OntoNotes 5.0 test set is comparable with that of other embeddings.

4.9 Ablation Study

To quantitatively and qualitatively analyze the effect of neutralization and repulsion in RAN-Debias, we perform an ablation study. We examine the following changes in RAN-Debias independently:

1. Nullify the effect of repulsion by setting $\lambda_1 = 0$, thus creating AN-GloVe.
2. Nullify the effect of neutralization by setting $\lambda_3 = 0$, thus creating RA-GloVe.

We demonstrate the effect of the absence of neutralization or repulsion through a comparative analysis on GIPE and the SemBias analogy test.

The GIPE values for AN-GloVe, RA-GloVe, and RAN-GloVe are presented in Table 8. We observe that in the absence of repulsion (AN-GloVe), the performance is degraded by at least 72% compared to RAN-GloVe. It indicates the efficacy of repulsion in our objective function as a way to reduce the unwanted gender-biased associations between words and their neighbors, thereby reducing GIPE. Further, even in the absence of neutralization (RA-GloVe), GIPE is worse by at least 50% as compared to RAN-GloVe. In fact, the minimum GIPE is observed for RAN-GloVe, where both repulsion and neutralization are used in synergy as compared to the absence of any one of them.

To illustrate further, Table 9 shows the rank of neighbors having illicit proximities for three professions, using different version of debiased embeddings. It can be observed that the ranks in RA-GloVe are either close to or further away from the ranks in AN-GloVe, highlighting the importance of repulsion in the objective function. Further, the ranks in RAN-GloVe are the farthest,

Word	**Neighbor**	**Embedding**		
		AN-GloVe	RA-GloVe	RAN-GloVe
Captain	sir	28	22	52
	james	26	30	75
Nurse	women	57	56	97
	mother	49	74	144
Farmer	father	22	54	86
	son	45	90	162

Table 9: For three professions, we compare the ranks of their neighbors due to illicit proximities (the values denote the ranks).

Dataset	**Embedding**	**Definition** ↑	**Stereotype** ↓	**None** ↓
SemBias	AN-GloVe	**93.0**	**0.2**	6.8
	RA-GloVe	83.2	7.3	9.5
	RAN-GloVe	92.8	1.1	**6.1**

Table 10: Comparison for the gender relational analogy test on the SemBias dataset. ↑ (↓) indicates that higher (lower) value is better.

corroborating the minimum value of GIPE as observed in Table 8.

Table 10 shows that in the absence of neutralization (RA-GloVe), the tendency of favouring stereotypical analogies increases by an absolute difference of 6.2% as compared to RAN-GloVe. On the other hand, through the presence of neutralization, AN-GloVe does not favor stereotypical analogies. This suggests that reducing the projection of biased words on gender direction through neutralization is an effective measure to reduce stereotypical analogies within the embedding space. For example, consider the following instance of word pairs from the SemBias dataset: {(*widower, widow*), (*book, magazine*), (*dog, cat*), (*doctor, nurse*)}, where *(widower, widow)* is a gender-definition word pair while *(doctor, nurse)* is a gender-stereotype word pair and the remaining are of none type as explained in Section 4.4. During the evaluation, RA-GloVe incorrectly selects the gender-stereotype word pair as the closest analogy with *(he, she)*, while AN-GloVe and RAN-GloVe correctly select the gender-definition word pair. Further, we observe that RAN-GloVe is able to maintain the high performance of AN-GloVe, and the difference is less (0.2% compared to 1.1%) which is compensated by the superior performance of RAN-GloVe over other metrics like GIPE.

Through this ablation study, we understand the importance of repulsion and neutralization in the multi-objective optimization function of RAN-Debias. The superior performance of RAN-GloVe can be attributed to the synergistic interplay of repulsion and neutralization. Hence, in RAN-GloVe we attain the best of both worlds.

4.10 Case Study: Neighborhood of Words

Here we highlight the changes in the neighborhood (collection of words sorted in the descending order of cosine similarity with the given word) of words before and after the debiasing process. To maintain readability while also demonstrating the changes in proximity, we only analyze a few selected words. However, our proposed metric GIPE quantifies this for an exhaustive vocabulary set.

We select a set of gender-neutral professions having high values of gender-based proximity bias η_{w_i} as defined earlier. For each of these professions, in Table 11, we select a set of four words from their neighborhood for two classes:

- **Class A**: Neighbors arising due to gender-based illicit proximities.
- **Class B**: Neighbors whose proximities are not due to any kind of bias.

For the words in class A, the debiasing procedure is expected to increase their rank, thereby decreasing the semantic similarity, while for words belonging to class B, debiasing procedure is expected to retain or improve the rank for maintaining the semantic information.

We observe that RAN-GloVe not only maintains the semantic information by keeping the

Word	Class	Neighbor	Embedding				
			GloVe	Hard-GloVe	GN-GloVe	GP-GloVe	RAN-GloVe
Captain	A	sir	19	32	34	20	52
		james	20	22	26	18	75
		brother	34	83	98	39	323
		father	39	52	117	40	326
	B	lieutenant	1	1	1	1	1
		colonel	2	2	2	2	2
		commander	3	3	4	3	3
		officer	4	5	10	4	15
Nurse	A	woman	25	144	237	16	97
		mother	27	71	127	25	144
		housekeeper	29	54	28	29	152
		girlfriend	32	74	60	31	178
	B	nurses	1	1	1	1	1
		midwife	2	3	2	3	2
		nursing	3	2	3	2	9
		practitioner	4	5	4	5	3
Socialite	A	businesswoman	1	1	1	1	6
		heiress	2	2	2	2	9
		niece	12	18	14	17	78
		actress	19	16	38	14	120
	B	philanthropist	3	3	3	3	1
		aristocrat	4	4	4	4	3
		wealthy	5	5	7	5	4
		socialites	6	15	5	9	10
Farmer	A	father	12	28	37	13	84
		son	21	84	77	26	162
		boy	50	67	115	45	105
		man	51	50	146	60	212
	B	rancher	1	2	1	2	3
		farmers	2	1	4	1	1
		farm	3	3	5	4	2
		landowner	4	4	2	5	5

Table 11: For four professions, we compare the ranks of their class A and class B neighbors with respect to each embedding. Here, rank represents the position in the neighborhood of a profession, and is shown by the values under each embedding.

rank of words in class B close to their initial value but unlike other debiased embeddings, it drastically increases the rank of words belonging to class A. However, in some cases like the word ‘‘Socialite’’, we observe that the ranks of words such as ‘‘businesswoman’’ and ‘‘heiress’’, despite belonging to class A, are close to their initial values. This can be attributed to the high semantic dependence of ‘‘Socialite’’ on these words, resulting in a bias removal and semantic information tradeoff.

5 Conclusion

In this paper, we proposed a post-processing gender debiasing method called RAN-Debias. Our method not only mitigates direct bias of a word but also reduces its associations with other words that arise from gender-based predilections. We also proposed a word classification method, called KBC, for identifying the set of words to be debiased. Instead of using ‘‘biased’’ word embeddings, KBC uses multiple knowledge bases for word classification. Moreover, we proposed Gender-based Illicit Proximity Estimate (GIPE), a metric to quantify the extent of illicit proximities in an embedding. RAN-Debias significantly outperformed other debiasing methods on a suite of evaluation metrics, along with the downstream application task of coreference resolution while introducing minimal semantic disturbance.

In the future, we would like to enhance KBC by utilizing machine learning methods to account for the words which are absent in the knowledge base. Currently, RAN-Debias is directly applicable to non-contextual word embeddings for non-gendered grammatical languages. In the wake of recent work such as Zhao et al. (2019), we would like to extend our work towards contextualized embedding models and other languages with grammatical gender like French and Spanish.

Acknowledgment

The work was partially supported by the Ramanujan Fellowship, DST (ECR/2017/00l691). T. Chakraborty would like to acknowledge the support of the Infosys Center for AI, IIIT-Delhi.

References

Abdulrahman Almuhareb and Massimo Poesio. 2005. Concept learning and categorization from the Web. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 27, pages 103–108.

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*, pages 4349–4357.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *CoRR*, abs/1904.03035.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

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. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275. Hong Kong, China. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716.

Stanisław Jastrzebski, Damian Leśniak, and Wojciech Marian Czarnecki. 2017. How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks. *arXiv preprint arXiv:1702.02170*.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR*, pages 1–15.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Thomas Manzini, Lim Yao Chong, Alan W. Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19: 313–330.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing lstm language models. *International Conference on Learning Representations*, pages 1–13.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013,Workshop Track Proceedings*, pages 1–12.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

George A. Miller. 1995. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Nikola Mrkšić, Diarmuid Ó. Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó. Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989.

Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6620–6631.

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.

Jane Pilcher. 2017. Names and ''doing gender'': How forenames and surnames contribute to gender identities, difference, and inequalities. *Sex Roles*, 77(11–12):812–822.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, pages 337–346. ACM.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).*

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana. Association for Computational Linguistics.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel-aati Hawwary, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2012. Ontonotes release 5.0.

Adina Williams, Damian Blasi, Lawrence Wolf-Sonkin, Hanna Wallach, and Ryan Cotterell. 2019. Quantifying the semantic core of gender systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5734–5739.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284.