Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

1 Introduction

There have been hundreds or thousands of papers written about word embeddings and their applications, from Web search [27] to parsing Curriculum Vitae [16]. However, none of these papers have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems.

A word embedding that represent each word (or common phrase) w as a d-dimensional *word vector* $\vec{w} \in \mathbb{R}^d$. Word embeddings, trained only on word co-occurrence in text corpora, serve as a dictionary of sorts for computer programs that would like to use word meaning. First, words with similar semantic meanings tend to have vectors that are close together. Second, the vector differences between words in embeddings have been shown to represent relationships between words [32, 26]. For example given an analogy puzzle, "man is to king as woman is to x" (denoted as *man*:*king* :: *woman*:x), simple arithmetic of the embedding vectors finds that x=*queen* is the best answer because:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

Similarly, x=*Japan* is returned for *Paris*:*France* :: *Tokyo*:x. It is surprising that a simple vector arithmetic can simultaneously capture a variety of relationships. It has also excited practitioners because such a tool could be useful across applications involving natural language. Indeed, they are being studied and used in a variety of downstream applications (e.g., document ranking [27], sentiment analysis [18], and question retrieval [22]).

However, the embeddings also pinpoint sexism implicit in text. For instance, it is also the case that:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Figure 1: The most extreme occupations as projected on to the *she−he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 2: **Analogy examples**. Examples of automatically generated analogies for the pair *she-he* using the procedure described in text. For example, the first analogy is interpreted as *she*:*sewing* :: *he*:*carpentry* in the original w2vNEWS embedding. Each automatically generated analogy is evaluated by 10 crowd-workers are to whether or not it reflects gender stereotype. Top: illustrative gender stereotypic analogies automatically generated from w2vNEWS, as rated by at least 5 of the 10 crowd-workers. Bottom: illustrative generated gender-appropriate analogies.

***softball* extreme**	**gender portion**	**after debiasing**
1. pitcher	-1%	1. pitcher
2. bookkeeper	20%	2. infielder
3. receptionist	67%	3. major leaguer
4. registered nurse	29%	4. bookkeeper
5. waitress	35%	5. investigator
***football* extreme**	**gender portion**	**after debiasing**
1. footballer	2%	1. footballer
2. businessman	31%	2. cleric
3. pundit	10%	3. vice chancellor
4. maestro	42%	4. lecturer
5. cleric	2%	5. midfielder

Figure 3: **Example of indirect bias**. The five most extreme occupations on the *softball-football* axis, which indirectly captures gender bias. For each occupation, the degree to which the association represents a gender bias is shown, as described in Section 5.3.

In other words, the same system that solved the above reasonable analogies will offensively answer "man is to computer programmer as woman is to x" with x=*homemaker*. Similarly, it outputs that a *father* is to a *doctor* as a *mother* is to a *nurse*. The primary embedding studied in this paper is the popular publicly-available word2vec [24, 25] embedding trained on a corpus of Google News texts consisting of 3 million English words and terms into 300 dimensions, which we refer to here as the w2vNEWS. One might have hoped that the Google News embedding would exhibit little gender bias because many of its authors are professional journalists. We also analyze other publicly available embeddings trained via other algorithms and find similar biases.

In this paper, we will quantitatively demonstrate that word-embeddings contain biases in their geometry that reflect gender stereotypes present in broader society. Due to their wide-spread usage as basic features, word embeddings not only reflect such stereotypes but can also amplify them. This poses a significant risk and challenge for machine learning and its applications.

To illustrate bias amplification, consider bias present in the task of retrieving relevant web pages for a given query. In web search, one recent project has shown that, when carefully combined with existing approaches, word vectors have the potential to improve web page relevance results [27]. As an example, suppose the search query is *cmu computer science phd student* for a computer science Ph.D. student at Carnegie Mellon University. Now, the directory[1] offers 127 nearly identical web pages for students — these pages differ only in the names of the students. A word embedding's semantic knowledge can improve relevance by identifying, for examples, that the terms *graduate research assistant* and *phd student* are related. However, word embeddings also rank terms related to computer science closer to male names than female names (e.g., the embeddings give *John*:*computer programmer* :: *Mary*:*homemaker*). The consequence is that, between two pages that differ only in the names *Mary* and *John*, the word embedding would influence the search engine to rank John's web page higher than Mary. In this hypothetical example, the usage of word embedding makes it even harder for women to be recognized as computer scientists and would contribute to widening the existing gender gap in computer science. While we focus on gender bias, specifically Female-Male (F-M) bias, the approach may be applied to other types of bias.

Uncovering gender stereotypes from text may seem like a trivial matter of counting pairs of words that occur together. However, such counts are often misleading [14]. For instance, the term *male nurse* is several times more frequent than *female nurse* (similarly *female quarterback* is many times more frequent than *male quarterback*). Hence, extracting associations from text, F-M or otherwise, is not simple, and "first-order" approaches would predict that the word *nurse* is more male than *quarterback*. More generally, Gordon and Van Durme show how *reporting bias* [14], including the fact that common assumptions are often left unsaid, poses a challenge to extracting knowledge from raw text. Nonetheless, $\overrightarrow{\text{nurse}}$ is closer to $\overrightarrow{\text{female}}$ than $\overrightarrow{\text{male}}$, suggesting that word embeddings may be capable of circumventing reporting bias in some cases. This happens because word embeddings are trained using second-order methods which require large amounts of data to extract associations and relationships about words.

The analogies generated from these embeddings spell out the bias implicit in the data on which they were trained. Hence, word embeddings may serve as a means to extract implicit gender associations from a large text corpus similar to how Implicit Association Tests [15] detect automatic gender associations possessed by people, which often do not align with self reports.

To quantify bias, we compare a word embedding to the embeddings of a pair of gender-specific words. For instance, the fact that $\overrightarrow{\text{nurse}}$ is close to $\overrightarrow{\text{woman}}$ is not in itself necessarily biased (it is also somewhat close to $\overrightarrow{\text{man}}$ – all are humans), but the fact that these distances are unequal suggests bias. To make this rigorous, consider the distinction between *gender specific* words that are associated with a gender by definition, and the remaining *gender neutral* words. Standard examples of gender specific words include *brother*, *sister*, *businessman* and *businesswoman*. The fact that $\overrightarrow{\text{brother}}$ is closer to $\overrightarrow{\text{man}}$ than to $\overrightarrow{\text{woman}}$ is expected since they share the definitive feature of relating to males. We will use the gender specific words to learn a gender subspace in the embedding, and our debiasing algorithm removes the bias only from the gender neutral words while respecting the definitions of these gender specific words.

We refer to this type of bias, where there is an association between a gender neutral word and a clear

[1] Graduate Research Assistants listed at `http://cs.cmu.edu/directory/csd`.

gender pair as *direct bias*. We also consider a notion of *indirect bias*,[2] which manifests as associations between gender neutral words that are clearly arising from gender. For instance, the fact that the word *receptionist* is much closer to *softball* than *football* may arise from female associations with both *receptionist* and *softball*. Note that many pairs of male-biased (or female-biased) words have legitimate associations having nothing to do with gender. For instance, while the words *mathematician* and *geometry* both have a strong male bias, their similarity is justified by factors other than gender. More often than not, associations are combinations of gender and other factors that can be difficult to disentangle. Nonetheless, we can use the geometry of the word embedding to determine the degree to which those associations are based on gender.

Aligning biases with stereotypes. Stereotypes are biases that are widely held among a group of people. We show that the biases in the word embedding are in fact closely aligned with social conception of gender stereotype, as evaluated by U.S.-based crowd workers on Amazon's Mechanical Turk.[3] The crowd agreed that the biases reflected both in the location of vectors (e.g. $\overrightarrow{\text{doctor}}$ closer to $\overrightarrow{\text{man}}$ than to $\overrightarrow{\text{woman}}$) as well as in analogies (e.g., *he*:*coward* :: *she*:*whore*) exhibit common gender stereotypes.

Debiasing. Our goal is to reduce gender biases in the word embedding while preserving the useful properties of the embedding. Surprisingly, not only does the embedding capture bias, but it also contains sufficient information to reduce this bias, as illustrated in 7. We will leverage the fact that there exists a low dimensional subspace in the embedding that empirically captures much of the gender bias. The goals of debiasing are:

1. Reduce bias:
   (a) Ensure that gender neutral words such as *nurse* are equidistant between gender pairs such as *he* and *she*.
   (b) Reduce gender associations that pervade the embedding even among gender neutral words.
2. Maintain embedding utility:
   (a) Maintain meaningful non-gender-related associations between gender neutral words, including associations within stereotypical categories of words such as fashion-related words or words associated with football.
   (b) Correctly maintain definitional gender associations such as between *man* and *father*.

Paper outline. After discussing related literature, we give preliminaries necessary for understanding the paper in Section 3. Next we propose methods to identify the gender bias of an embedding and show that w2vNEWS exhibits bias which is aligned with common gender stereotypes (Section 4). In Section 5, we define several simple geometric properties associated with bias, and in particular discuss how to identify the gender subspace. Using these geometric properties, we introduce debiasing algorithms (Section 6) and demonstrate their performance (Section 8). Finally we conclude with additional discussions of related literature, other types of biases in the embedding and future works.

2 Related work

Related work can be divided into relevant literature on bias in language and bias in algorithms.

[2]The terminology indirect bias follows Pedreshi et al. [29] who distinguish *direct* versus *indirect* discrimination in rules of fair *classifiers*. Direct discrimination involves directly using sensitive features such as gender or race, whereas indirect discrimination involves using correlates that are not inherently based on sensitive features but that, intentionally or unintentionally, lead to disproportionate treatment nonetheless.

[3]`http://mturk.com`