AI Ethics

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

M. Mehdi Balouchi
Prof. Flek

Definitions

Stereotypes: Imply placing a piece of information into a certain category based on the type of data it has stored (Brahnam & De Angeli, 2012)

Gender stereotypes: Are both descriptive, meaning that they are formed around a characteristic a man or a woman possesses, and prescriptive, which reflects the social perception of what a person should be according to their gender (Brahnam & De Angeli, 2012).

Stereotypes are so powerful that not only are they generated in the human-to-human interactions, they are also applied to non-human entities (Brahnam & De Angeli, 2012). The relation between technology and gender stereotypes started to become a subject of interest in the 1990's, when the presence of stereotypes was identified in interactions with computers (Nass, Moon, & Green, 1997).

Examples:

Early Years

Girls should play with dolls and boys should play with trucks

Boys should be directed to like blue and green; girls toward red and pink

Boys should not wear dresses or other clothes typically associated with "girl's clothes"

During Youth

Girls are better at reading and boys are better at math

Girls should be well behaved; boys are expected to act out

Girls and are not as interested as boys in STEM subjects

Boys should engage in sports and refrain from more creative pursuits

Boys and men are expected to use violence and aggression to prove their manliness

A boy that doesnt use violence or aggression is an understandable target for bullying

Girls should be thin and beautiful to make them appealing to men

As Adults

Victims of intimate partner violence are weak because they stay in the relationship

There is something wrong with a woman who doesn’t want children

Assertive women are unfeminine and are "bossy," "bitches” or “whores”

Women are natural nurturers; men are natural leaders

Women don’t need equal pay because they are supported by their husbands

Women with children are less devoted to their jobs

Men who spend time with family are less masculine and poor breadwinners

In heterosexual couples, women should take time off to care for children or elders

Men who are not aggressive and/or assertive are unmanly and likely gay

Women are too emotional to undertake certain kinds of work, especially while pregnant

Men are too impersonal and not emotionally apt to take on tasks "better done by women"

Gender Bias In AI

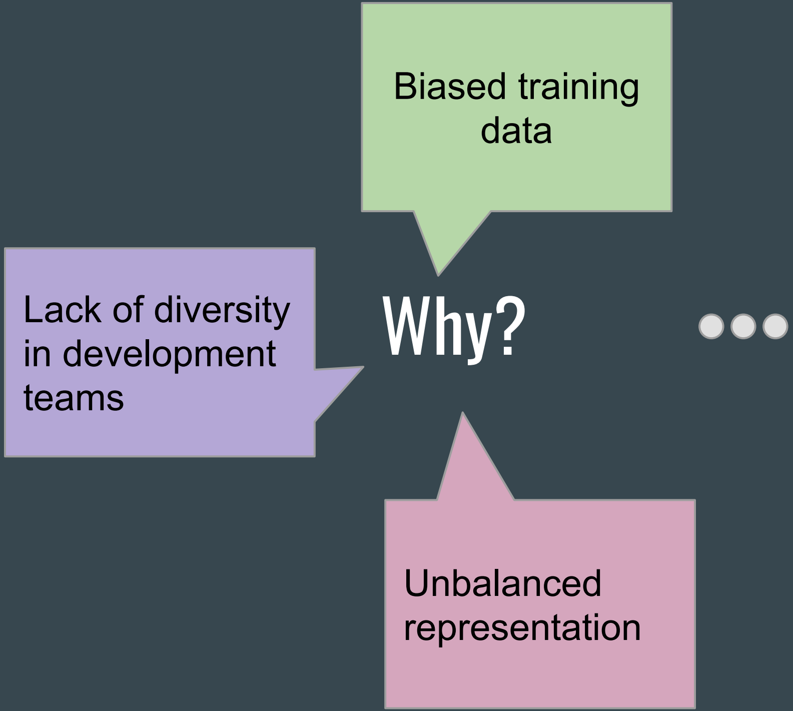
Hiring bias

Facial recognition

Virtual assistants

Healthcare bias

Language Models



Why?

Biased training data

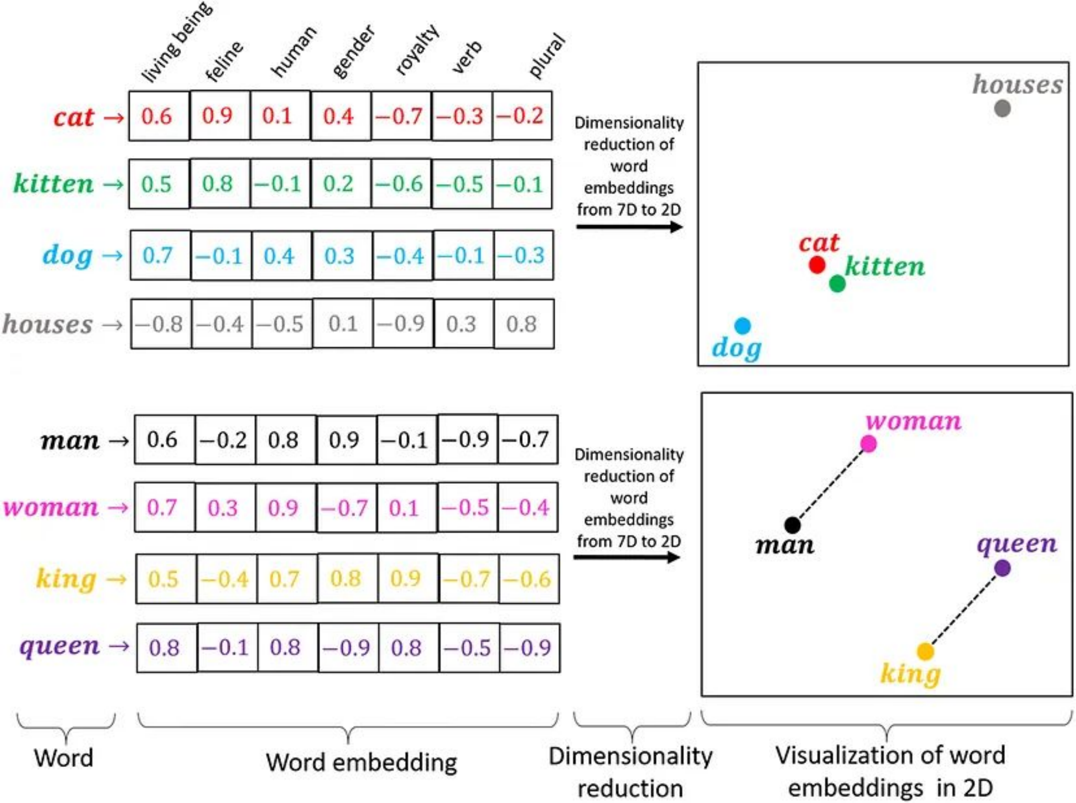
Lack of diversity in development teams

Unbalanced representation

●●●

Word Embeddings

1. Words => High-dimensional vector space
2. Words that have similar meanings are located close to each other
3. Words that are dissimilar are located farther Apart
4. **The vector differences between words in embeddings have been shown to represent relationships between words**



	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Word Embeddings

Word2Vec: Word2Vec is one of the most popular algorithms for generating word embeddings. It has been used in a wide range of natural language processing (NLP) applications, such as sentiment analysis, named entity recognition, and machine translation.

GloVe: GloVe (Global Vectors for Word Representation) is another popular algorithm for generating word embeddings. It uses a co-occurrence matrix to learn vector representations of words, with the goal of minimizing the difference between the dot product of two word vectors and the logarithm of their co-occurrence count.

fastText: fastText is a word embedding algorithm developed by Facebook that is based on Word2Vec, but incorporates subword information. It is designed to handle out-of-vocabulary words and can generate word embeddings for rare words or misspellings.

ELMo: ELMo (Embeddings from Language Models) is a deep contextualized word embedding model that was introduced by researchers at Allen Institute for Artificial Intelligence (AI2). It generates word embeddings by taking into account the context in which the word appears, and has been used in a variety of NLP tasks, such as question answering and sentiment analysis.

BERT: BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that generates contextualized word embeddings by training on large amounts of text data. It has achieved state-of-the-art performance on a wide range of NLP tasks, such as natural language inference, sentiment analysis, and question answering.

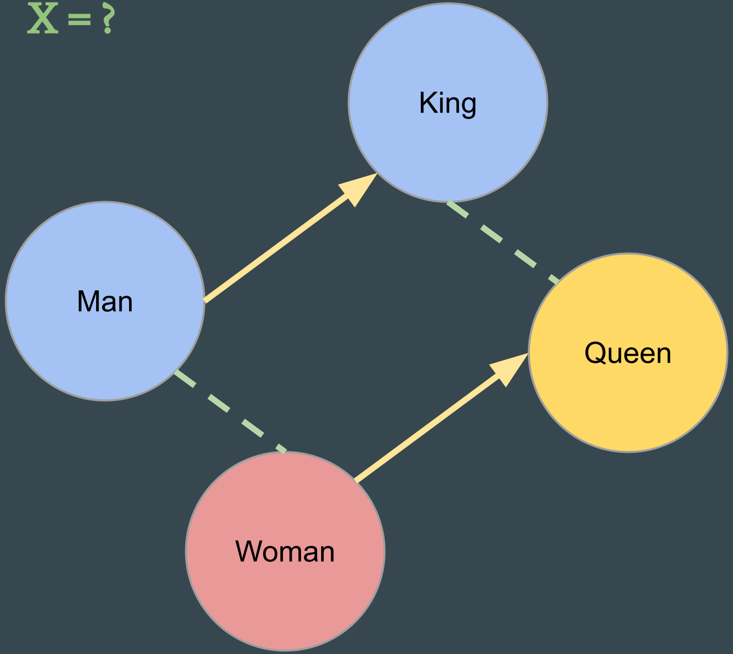
Word Embeddings - w2vNEWS

- Word2vec was created, patented and published in 2013 by a team of researchers led by Tomas Mikolov at Google
- Publicly-available
- 300 dimensional embedding
- 50,000 most frequent words =>After filtering, 26,377 words remained
- trained on a corpus of Google News texts
- 3 million English words
- Pretrained-Model accessible at https://code.google.com/archive/p/word2vec/

Analogy Puzzle

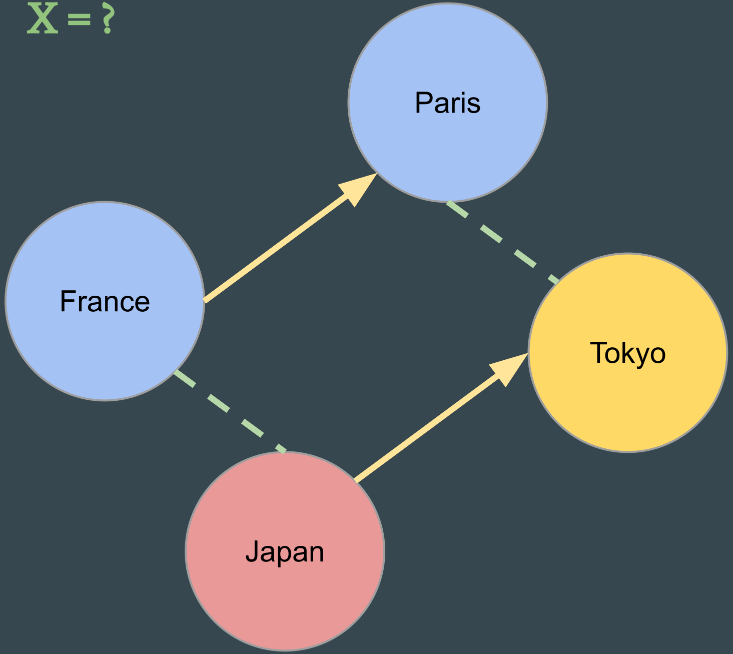
Man is to king as woman is to x

X = ?



France is to Paris as Japan is to x

X = ?



Analogy Puzzle - Sexism

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

$$\overrightarrow{\text{father}} - \overrightarrow{\text{mother}} \approx \overrightarrow{\text{doctor}} - \overrightarrow{\text{nurse}}$$

Gender appropriate she-he analogies - Gender Specific Words

she	he
1. Queen	1. King
2. Sister	2. Brother
3. Mother	3. Father
4. Waitress	4. Waiter
5. **Businesswoman**	5. **Businessman**

Gender Biased Occupations

Extreme she	Extreme he
1. Homemaker	1. Maestro
2. Nurse	2. Skipper
3. Receptionist	3. Protege
4. Librarian	4. Philosopher
5. Socialite	5. Captain
6. Hairdresser	6. Architect
7. nanny	7. Financier
8. Bookkeeper	8. Warrior
9. Stylist	9. Broadcaster
10. Housekeeper	10. Magician

List of the occupations that are closest to she and to he in the w2vNEWS embeddings.

Problem? Does it really matter?

Since word embeddings are often used as basic features in downstream NLP tasks such as sentiment analysis, machine translation, or speech recognition, these biases and stereotypes can be amplified and further propagated by the algorithms and models that rely on them.

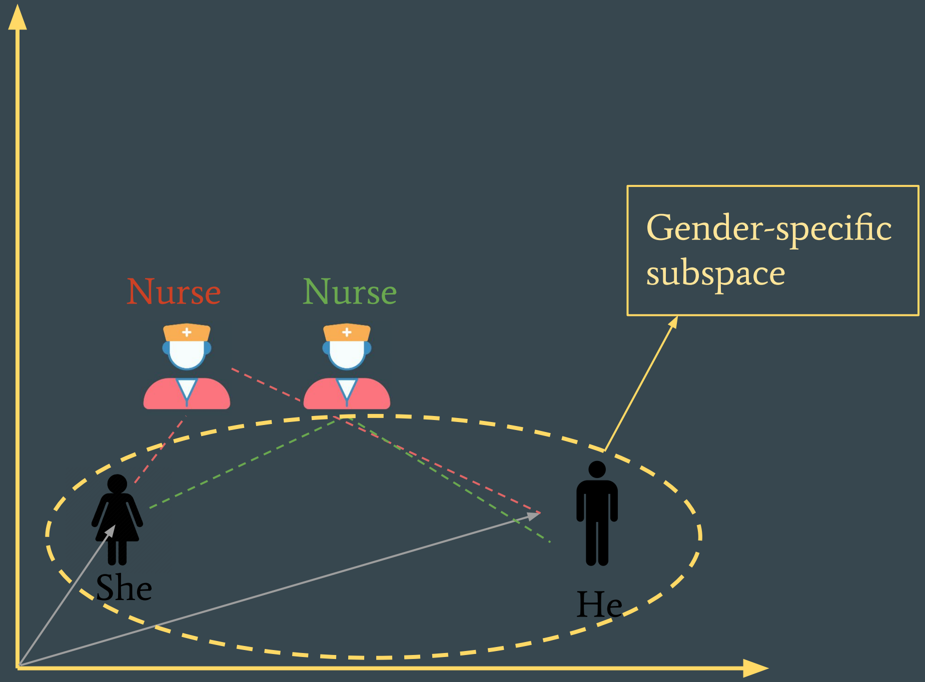
It is important for NLP researchers and practitioners to be aware of the potential biases in word embeddings and to take measures to mitigate them, such as debiasing techniques or using alternative representations of language that are less susceptible to such biases.

Gender bias and stereotype in Language (English) and Algorithms

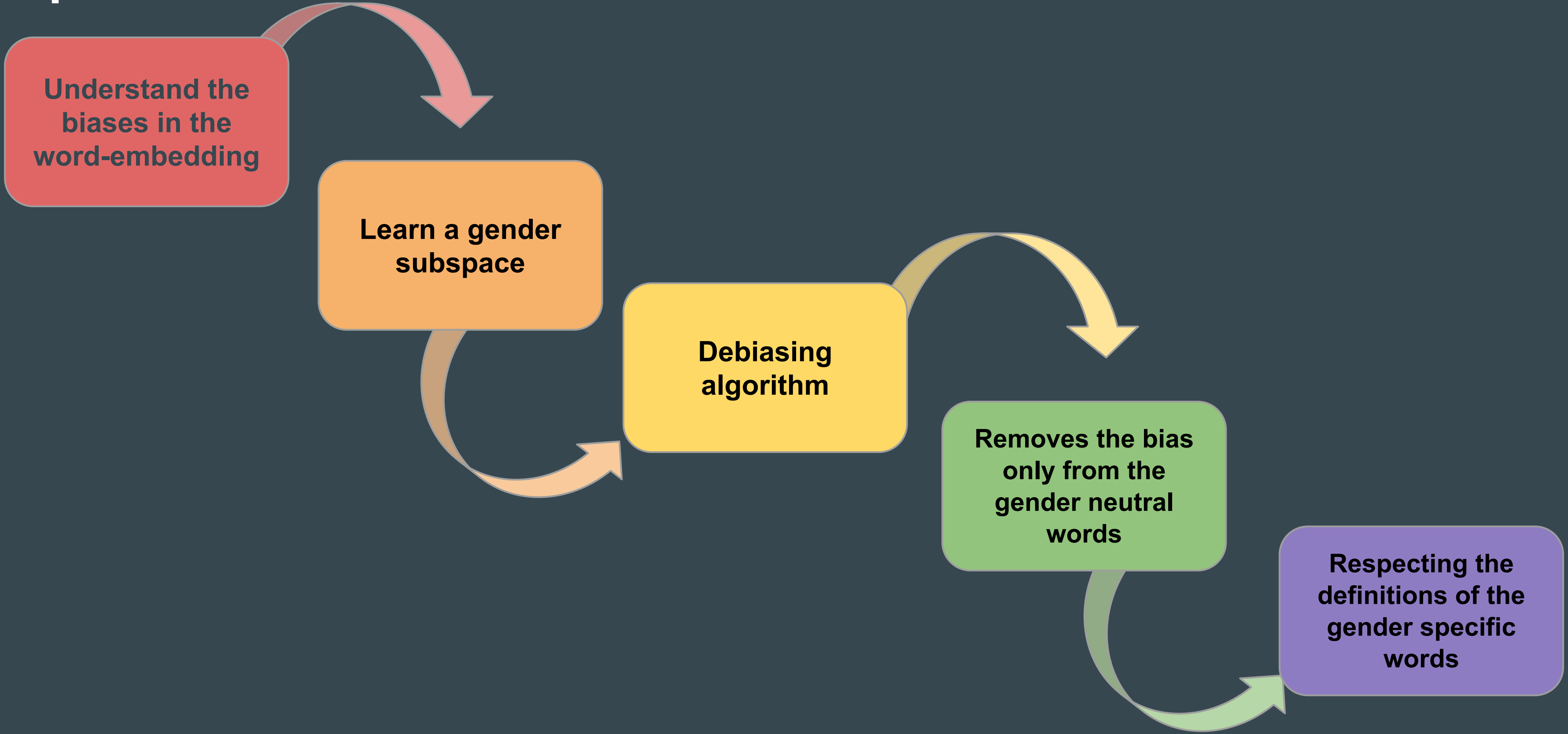
- It is important to quantify and understand bias in languages as such biases can reinforce the psychological status of different groups.
- Tests have uncovered gender-word biases that people do not self-report and may not even be aware of.
- A number of online systems have been shown to exhibit various biases, such as racial discrimination and gender bias in the ads presented to users

Quantifying The Bias

compare a word vector to the vectors of a pair of gender-specific words



Pipeline



Understand the biases in the word-embedding

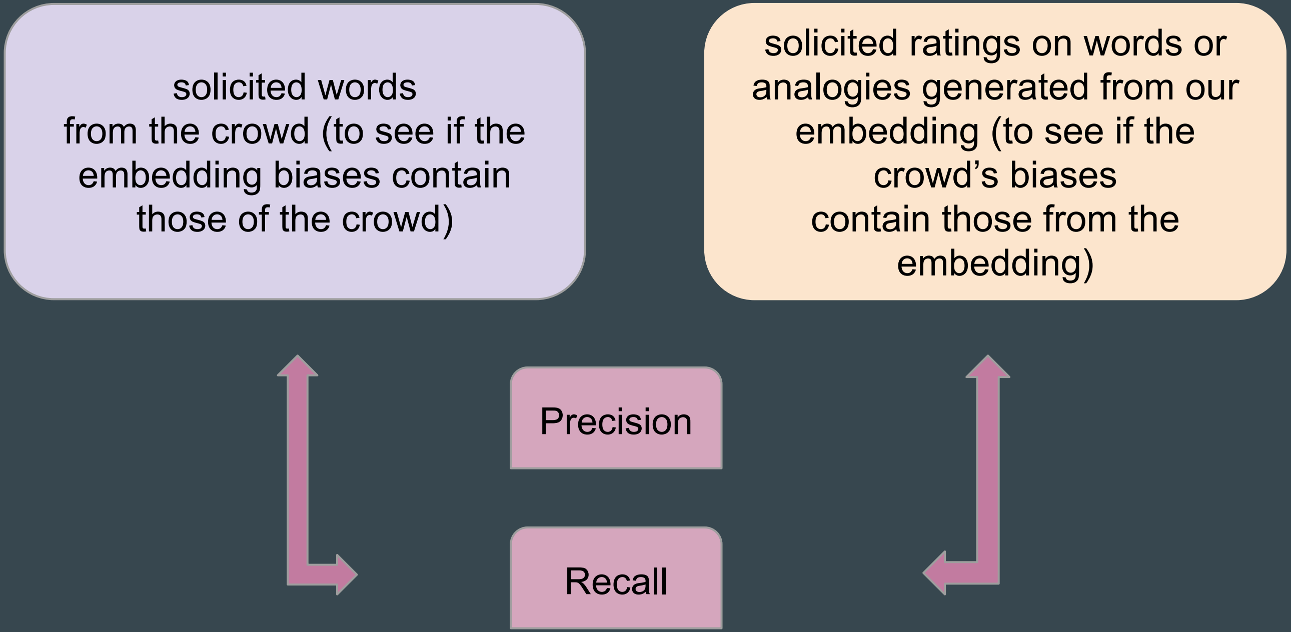
Learn a gender subspace

Debiasing algorithm

Removes the bias only from the gender neutral words

Respecting the definitions of the gender specific words

Crowd experiments - Amazon Mechanical Turk platform



solicited words from the crowd (to see if the embedding biases contain those of the crowd)

solicited ratings on words or analogies generated from our embedding (to see if the crowd’s biases contain those from the embedding)

Precision

Recall

1. Understand the biases present in the word-embedding

Occupational stereotypes

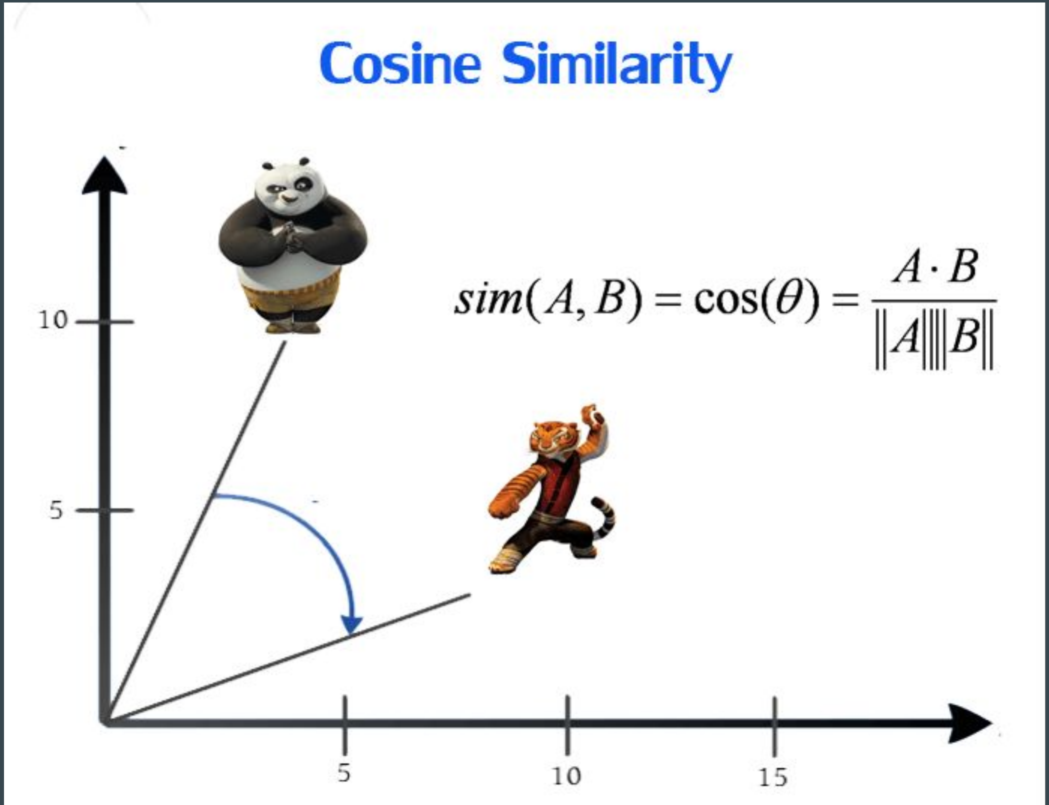
- lists the occupations that are closest to she and to he in the w2vNEWS embeddings
- Ask the crowdworkers to evaluate whether an occupation is considered female-stereotypic, male-stereotypic, or neutral

Analogies exhibiting stereotypes

- Analogies are a useful way to both evaluate the quality of a word embedding and also its stereotypes
- Use an algorithm to generate useful pairings
- Use crowdworkers to determine (a) whether the generated pairing makes sense as an analogy, and (b) whether it reflects a gender stereotype

Cosine Similarity

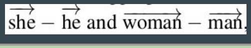
Cosine similarity is an important concept in NLP because it is a way to measure the similarity between two vectors that represent pieces of text.



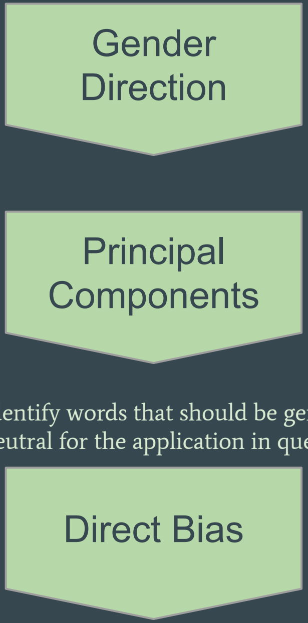
$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Ref: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

2. Identifying the gender subspace



$\overrightarrow{she} - \overrightarrow{he}$ and $\overrightarrow{woman} - \overrightarrow{man}$.

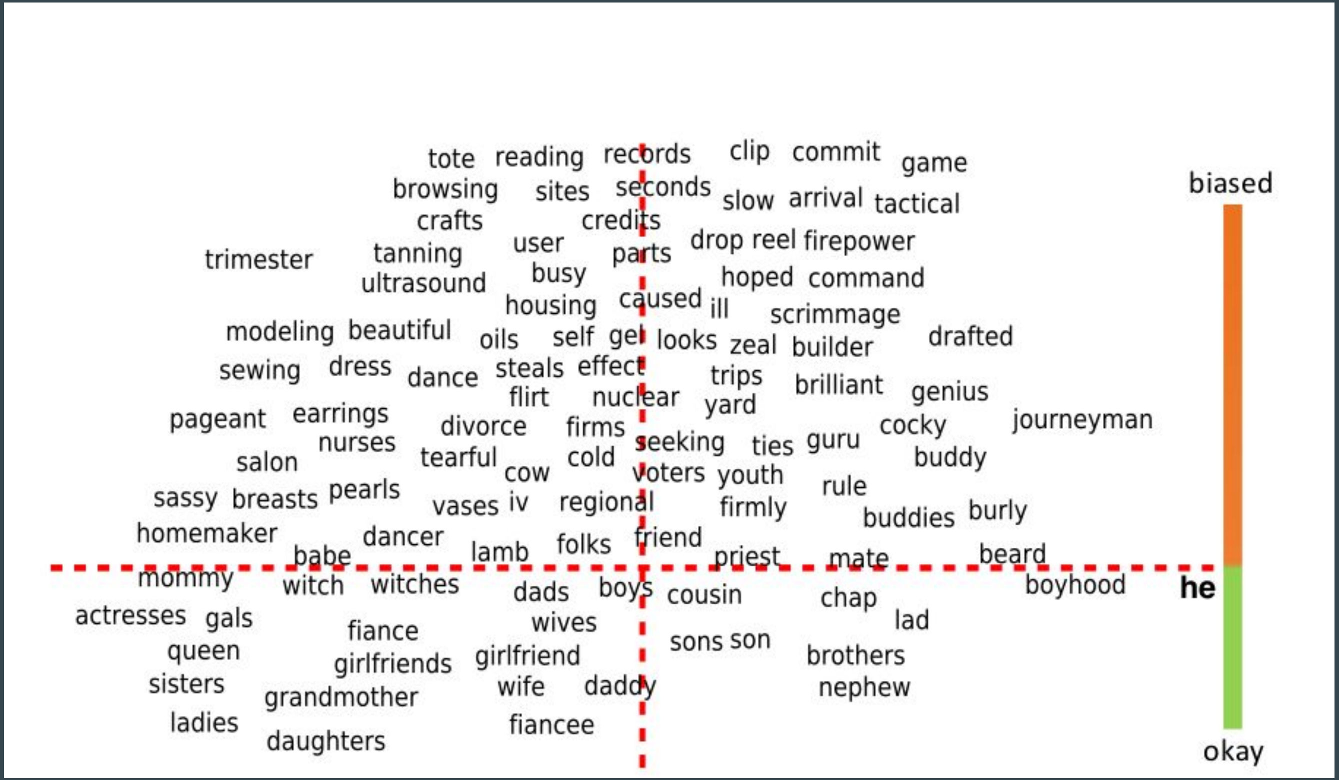


Gender Direction

Principal Components

Identify words that should be gender-neutral for the application in question

Direct Bias



3. Debiasing Algorithm

- The debiasing algorithms are defined in terms of sets of words rather than just pairs, so that we can consider other biases such as racial or religious biases

- We also assume that we have a set of words to neutralize

- Hard and Soft debiasing algorithms

4. Neutralize and Equalize - Soften

Neutralize and Equalize

- Ensures that gender neutral words are zero in the gender subspace
- The disadvantage of Equalize is that it removes certain distinctions that are valuable in certain applications.

Soften

- Ensures that any neutral word is equidistant to all words in each equality set

Debiasing Result

The result ensures that the word embeddings preserve the desirable properties of the original embedding while reducing both direct and indirect gender biases

1. Analogy Generation: automatically generated pairs of words that are analogous to she-he and asked crowd-workers to evaluate

   Example: He to Doctor => She to X, X was nurse and now is physician.

2. Check if new embeddings and the transformation does not negatively impact the performance by testing the debiased embedding on several standard benchmarks

Discussion

- word embeddings help us further our understanding of bias in language
- change the embeddings of gender neutral words, by removing their gender associations.
- Hard-debiasing algorithm significantly reduces both direct and indirect gender bias while preserving the utility of the embedding
- Soft-embedding algorithm which balances reducing bias with preserving the original distances, and could be appropriate in specific settings
- Debiasing in society > Debiasing in embeddings

References

1. Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29 (2016).
2. Leavy, Susan. "Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning." In Proceedings of the 1st international workshop on gender equality in software engineering, pp. 14-16. 2018.
3. Craiut, Miruna-Valeria, and Ioana Raluca Iancu. "Is technology gender neutral? A systematic literature review on gender stereotypes attached to artificial intelligence." Human Technology 18, no. 3 (2022): 297-315.
4. https://www.genderequalitylaw.org/examples-of-gender-stereotypes
5. Levy, Omer, and Yoav Goldberg. "Linguistic regularities in sparse and explicit word representations." In Proceedings of the eighteenth conference on computational natural language learning, pp. 171-180. 2014.