DREXEL UNIVERSITY College of Computing & Informatics

2022 Drexel Emerging Graduate Scholars Conference

Fairness in Word Embeddings: Definitions, Measurements, and Mitigations

REPORTER: Lu Wang

Advisor: Dr. Jina Huh-Yoo
PhD Program in Information Science
College of Computing and Informatics

OUTLINE

CONTENT

01 Introduction

Word embeddings, Social bias & debiasing in word embeddings, Insufficiency in existing methods, Validity

02 Methods

Research Questions, Data collection & Analysis frameworks

03 Results

Validity examinations on Definitions, Meas urements & Mitigations

04 Discussion

Contributions & Future work

PART ONE

01

Introduction

Word embeddings

Healthcare

Agmon, S., et al. (2022). "Gender-sensitive word embeddings for healthcare." Journal of the American Medical Informatics Association **29**(3): 415-423.

Recruitment

Qin, C., et al. (2018). Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. The 41st international ACM SIGIR conference on research & development in information retrieval.

Education

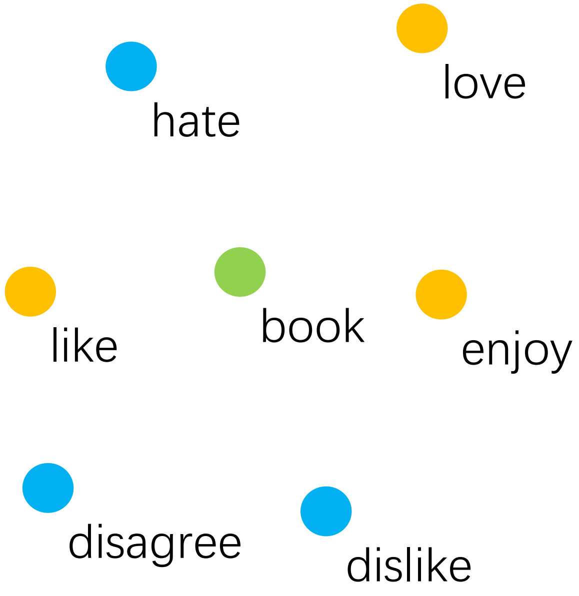
Li, H. and Y. Sun (2018). English education text recommendation technology based on word embedding. 2018 International Conference on Big Data and Artificial Intelligence (BDAI), IEEE.

Criminal Prediction

Zhang, Y., et al. (2020). "Predicting time and location of future crimes with recommendation methods." Knowledge-Based Systems **210**: 106503.

Word embeddings

Word embeddings map words into metric vectors and use the distance between the vectors to capture semantic information.



The objective of a Word2Vec model is to maximize the average log probability of each word's context following

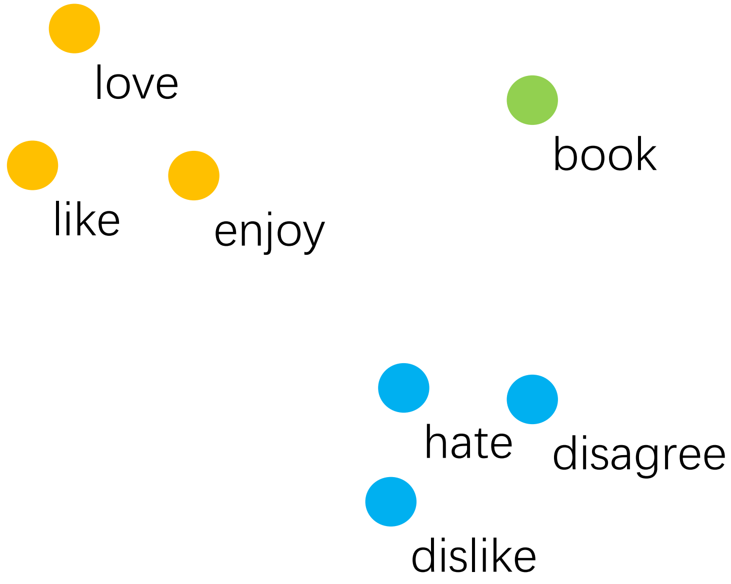
$$J = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log p(w_{t+j}|w_t), \qquad (1)$$

where T is the number of training words and c is the number of context words. $p(w_{t+j}|w_t)$ is given by the softmax function,

$$p(w_o|w_i) = \frac{\exp(v_{w_o}'^{\top} v_{w_i})}{\sum_{w=1}^{W}\exp(v_w'^{\top} v_{w_i})}, \qquad (2)$$

where W is the number of unique words (type) in the corpus $w_1 \dots w_T$, and v_w and v_w' are the input and output vector representations of word w.

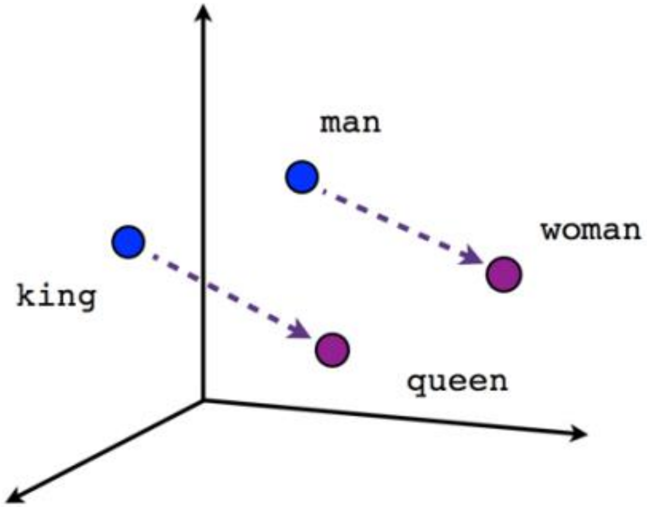
- Prediction-based Models
- Count-based Models



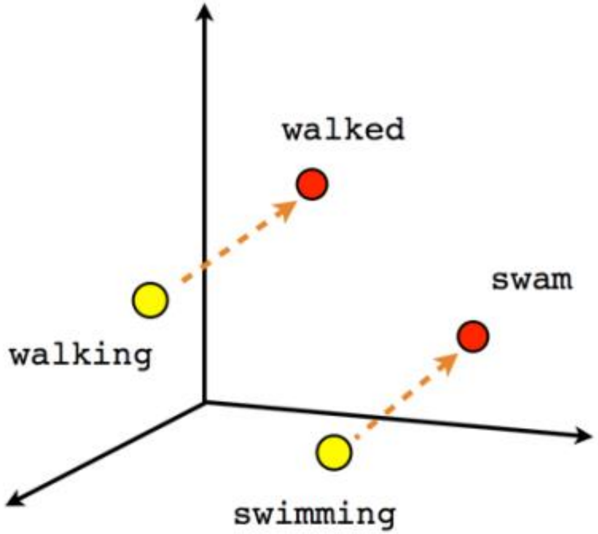
Almeida, F. and G. Xexéo (2019). "Word embeddings: A survey." arXiv preprint arXiv:1901.09069.

Word embeddings

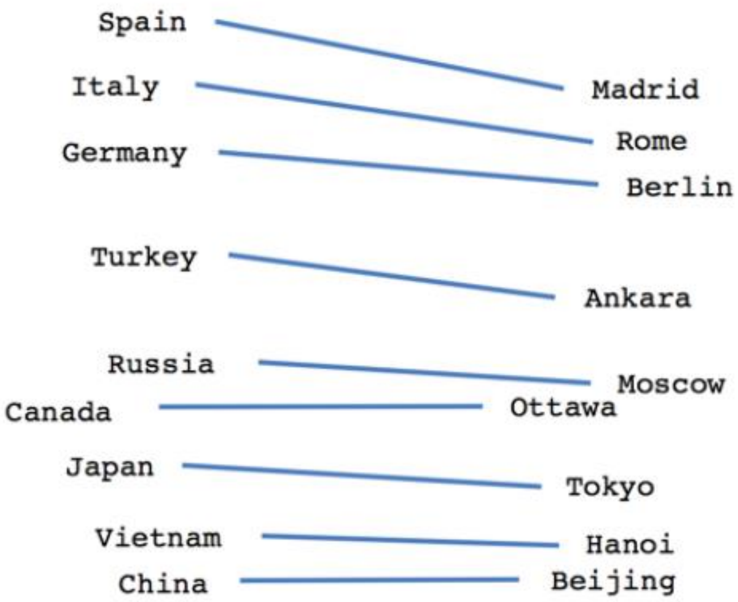
Embeddings can produce remarkable analogies.



Male-Female



Verb tense



Country-Capital

https://www.tensorflow.org/images/linear-relationships.png

Social bias in word embeddings

x=Japan is returned for Paris : France :: Tokyo : x

x=queen is returned for man : king :: woman : x

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

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

Bolukbasi, T., et al. (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29: 4349-4357.

Social bias in word embeddings

x=Japan is returned for Paris : France :: Tokyo : x

x=queen is returned for man : king :: woman : x

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

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

	Gender stereotype *she*-*he* analogies	
sewing-carpentry	registered 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	lovely-brilliant

Racial Analogies	
black → homeless	caucasian → servicemen
caucasian → hillbilly	asian → suburban
asian → laborer	black → landowner
Religious Analogies	
jew → greedy	muslim → powerless
christian → familial	muslim → warzone
muslim → uneducated	christian → intellectually

Table 1: Examples of racial and religious biases in analogies generated from word embeddings trained on the Reddit data from users from the USA.

Bolukbasi, T., et al. (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29: 4349-4357.
Manzini, T., et al. (2019). "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings." arXiv preprint arXiv:1904.04047.

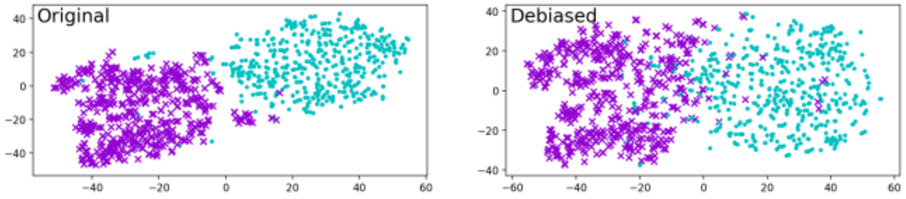
Debiasing in word embeddings

Figure 3: Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words *he* and *she*, and y is a direction learned in the embedding that captures gender neutrality, with gender neutral words above the line and gender specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line.

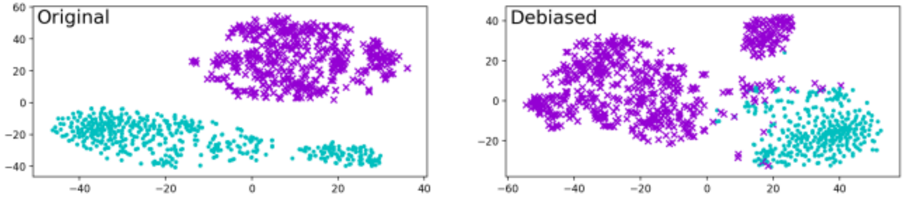
Bolukbasi, T., et al. (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29: 4349-4357.

Insufficient Debiasing in Word embeddings

Male- and female-biased words cluster together



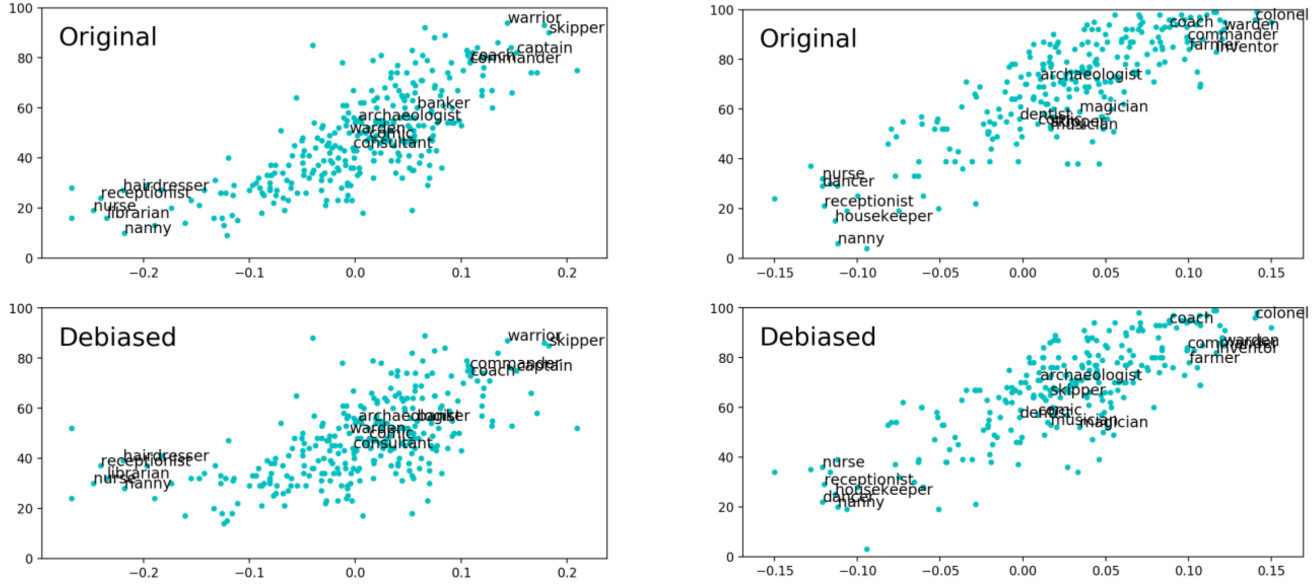
(a) Clustering for HARD-DEBIASED embedding, before (left hand-side) and after (right hand-side) debiasing.



(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

Figure 1: Clustering the 1,000 most biased words, before and after debiasing, for both models.

Bias-by-projection correlates to bias-by-neighbours



(a) The plots for HARD-DEBIASED embedding, before (top) and after (bottom) debiasing.

(b) The plots for GN-GLOVE embedding, before (top) and after (bottom) debiasing.

Figure 2: The number of male neighbors for each profession as a function of its original bias, before and after debiasing. We show only a limited number of professions on the plot to make it readable.

Gonen, H. and Y. Goldberg (2019). "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them." arXiv preprint arXiv:1903.03862.

Insufficient Debiasing in Word embeddings

Definitions

Philosophy & Theory level

Understanding & measuring social bias in word embeddings

Measurements

Specific methods and tasks to measure

Mitigations

To mitigate bias based on definitions & measurements

Measure

Insufficient Debiasing in Word embeddings

If mitigations are successful but insufficient, then the measurements are possibly insufficient.

Definitions

Philosophy & Theory level

Understanding & measuring social bias in word embeddings

Measurements

Specific methods and tasks to measure

Mitigations

To mitigate bias based on definitions & measurements

Measure

Insufficient Debiasing in Word embeddings

If mitigations are successful but insufficient, then the measurements are possibly insufficient.

Validity

"The extent to which a test measures what it purports to measure."

1921 by the National Association of the Directors of Educational Research

"Validity describes the extent to which a measure accurately represents the concept it claims to measure."

Punch, K. F. (2013). Introduction to social research: Quantitative and qualitative approaches, sage.

Urbina, S. (2014). Essentials of psychological testing, John Wiley & Sons.
Roberts, P. and H. Priest (2006). "Reliability and validity in research." Nursing standard 20(44): 41-46.

Validity for measurements

Validity

Internal

The reasons of the outcomes and reduce other unanticipated reasons

Content validity:

relevance & representativeness

[indexes or variables to measure]

Criterion-related validity:

can be compared to other similar validated measures of the same concept or phenomenon

[comparisons with other measurements]

Construct validity:

demonstrating relationships between the concepts and the construct or theory

[definitions]

External

Can be applied to other people and other situations.

[other word embeddings, other datasets]

Motivation

To understand the insufficiency of debiasing in word embeddings by

investigating the definitions, measurements, and mitigations in terms of validity.

02

PART TWO

Methods

Research Questions

This study is going to understand:

1. how researchers defined bias measurements in word embeddings
2. how researchers measured and evaluated the bias in word embeddings
3. how researchers mitigated and reduced the detected bias in word embeddings

Understanding these research questions will help to define new sufficient definitions.

Data collection

995 papers were collected by Herzing's Publish or Perish software (Harzing, 2007) from Google Scholar with search query of "(bias OR debias OR fairness) AND 'word embedding'" on January 30th, 2022. The publication years range from 2013 to 2022. Among them, 165 papers are with citations above the average of 48.17. Finally, ten papers were identified by the criteria of:

1) English word embeddings,
2) focusing on the fairness problem of word embeddings,
3) including methods to measure and evaluate the bias, and
4) including the methods to mitigate and reduce the bias.

Analysis frameworks

Definitions:

Content validity

what the labels were, what the sensitive attributes were,

Construct validity

how the researchers defined measurements

Measurements :

External validity

what kind of word embeddings the researchers used

Criterion-related validity

what kind of tasks the researchers applied to measure

Mitigations:

in what stage the mitigation involved and how researchers operated when they found bias in word embeddings

PART THREE

03

Results

Content validity: mostly limited to gender bias

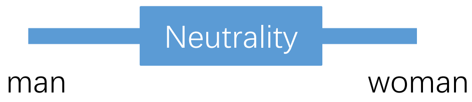
what the labels were, what the sensitive attributes were,

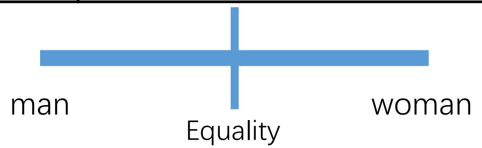
Construct validity: what is the actual construct of gender bias?

how the researchers defined measurements: three types

Definitions

Type	Paper	Labels	Attributes	Measuring bias in word embeddings
Neutrality	1	Selected analogies	Genders	Humans' ratings, Gender subspace: Using a threshold to decide whether the projection on subspace refers to bias.
Neutrality	2	Occupations	Genders	Minimization objective $J = JG + \lambda dJD + \lambda eJE$ Male-definition, female definition, and gender-neutral definitions by WordNet.
Neutrality	3	Bias datasets	Gender, race, religion	Bias subspace Using a threshold to decide whether the projection on subspace refers to bias.
Neutrality	4	Occupations	Genders	Humans' ratings, Gender subspace: Using a threshold to decide whether the projection on subspace refers to bias.
Neutrality	5	Bias datasets	Genders	Against several baselines: Better than baselines
Neutrality	6	Bias datasets	Genders	$L = \lambda fLf + \lambda mLm + \lambda gLg + \lambda rLr$, Male-definition, female definition, by WordNet, and gender-neutral definitions by annotators.
Equality	7	Occupations	Genders	The difference in word counts, visualization distance, accuracy, F1 scores The difference between pro/anti stereotype ($p < .05$) means bias.
Equality	8	Bias datasets	Genders, pleasant and unpleasant terms	WEAT: A significant difference between two target sets and two attributes sets means bias.
Neutrality, Equality	9	Names	Genders	Humans' ratings, WEAT: A significant difference between two target sets and two attributes sets means bias.
Mediator	10	Occupations	Genders & Names	Performance when using names and genders to debias: Similar results when using names and genders to debias.





names → Gender info

External validity: few tested on multiple word embeddings

what kind of word embeddings the researchers used

Criterion-related validity

what kind of tasks the researchers applied to measure

Measurements

Papers	Word embeddings	Measurements & Evaluations
1	Word2Vec	Standard evaluation metrics, U.S. based crowd-workers to evaluate the analogies
2	GloVe	Visualization, Gender Relational Analogy, Word Similarity, and Analogy, Coreference Resolution
7	ELMo, GloVe	Counts for the number of occurrences of male pronouns (he, his, and him) and female pronouns (she and her) in the corpus as well as the co-occurrence of occupation words with those pronouns, Principal components analysis, Visualize the gender subspace, Classifier accuracy, Bias in Coreference Resolution,
3	Word2Vec	Mean average cosine similarity (MAC), P-values to measure the effects of debiasing. Tasks of NER, POS tagging, and POS chunking
8	GloVe	The effect size of two different WEAT biases, Correlations with the ground truth change in bias (as measured by retraining the embedding after removing a subset of the training corpus).
4	Word2Vec	Stereotyped analogies, Amazon Mechanical Turk to evaluate the analogies, Variance of the projections in the original embedding and after the debiasing transformation, Absolute values of the projections onto the he-she direction before and after debasing, Tasks of Semantic Similarity Measurements, MSR-analogy
10	GloVe	WEAT, Analogies, Embedding Coherence Test (ECT), Embedding Quality Test, ECT (word pairs) uses E defined by gendered word pairs and ECT (names) which uses vectors m and s derived by gendered names, Cosine similarity on WordSimilarity 353 and SimLex-999, each of which evaluates a Spearman coefficient. Google Analogy Dataset using the function 3COSADD
9	Word2Vec, fast, GloVe	WEAT, US-based crowd workers on Amazon's Mechanical Turk, Potential indirect bias metrics
5	Word2Vec	WEAT, Cohen's d, One-sided p-values, Indirect bias, Cluster, Reclassify performance, Word similarity, Sentiment classification, Non-biased gender analogies (error rate)
6	GloVe, or any pre-trained word embeddings	SemBias & SemBias-subset test, Analogy Detection, Semantic Similarity Measurement, Visualizing the Effect of Debiasing.

External validity: few tested on multiple word embeddings

what kind of word embeddings the researchers used

Criterion-related validity: multiple measures to compare but only limited to pre-debiasing and post-debiasing. Lack the correlations of different measurements

what kind of tasks the researchers applied to measure

Measurements

Papers	Word embeddings
1	Word2Vec
2	GloVe
7	ELMo, GloVe
3	Word2Vec
8	GloVe
4	Word2Vec
10	GloVe
9	Word2Vec, fast, GloVe
5	Word2Vec
6	GloVe, or any pre-trained word embeddings

There are five types of measurements and evaluations:

1) established standard evaluations which provide test scores to compare,
2) statistical indicators like Cohen’s d and p-values,
3) downstream tasks like classification and coreference solutions,
4) human ratings, taken through Amazon Mechanical Turk platforms,
5) visualizations of analogies or clusters after debiasing.

Mitigations

The stages included pre-processing, in-processing, and post-processing. Two required re-train the word embeddings.
There are four types of mitigations,

1) removal of the biased data,
2) under equality, to create equal corpora for the attributes,
3) under neutrality, training word embeddings with objective functions,
4) to define subspace and manipulate the biased words in post-processing stage.

Papers	**Stages**	**Operations**
5	Pre-processing	Counterfactual Data, Substitution (CDS), Names Intervention, bipartite-graph matching of names by frequency and gender-specificity.
7	Pre- & Post-processing	Data augmentation approach, gender-swapped version of the test instances use their average as the final representations, A test-time neutralization approach.
10	Pre- & post-processing	Use names as an alternative to bootstrap finding the gender direction, Data subtraction Bias subspace
2	In-processing	Minimizing the negative distances between words in the two groups, to be retained in the null space of the gender direction, preserving gender information in certain dimensions of word vectors while compelling other dimensions to be free of gender influence.
6	In-processing	Four types of information: feminine, masculine, gender-neutral, and stereotypical, which represent the relationship between gender vs. bias, and propose a debiasing method that (a) preserves the gender-related information in feminine and masculine words, (b) preserves the neutrality in gender-neutral words, and (c) removes the biases from stereotypical words.
4	post-processing	Transformed embeddings are stereotypical-free, labels should be perpendicular to gender direction
1	Post-processing	Bias subspace
3	Post-processing	Bias subspace in a multiclass setting,
8	Post-processing & Pre-processing (re-train)	To identify subsets of documents whose removal would most reduce bias
9	Post-processing & Pre-processing (re-train)	Removing names

Mitigations

The stages included pre-processing, in-processing, and post-processing. Two required re-train the word embeddings.
There are four types of mitigations,

1) removal of the biased data,
2) under equality, to create equal corpora for the attributes,
3) under neutrality, training word embeddings with objective functions,
4) to define subspace and manipulate the biased words in post-processing stage.

Depending on sufficient measurements.

	Data removal	Data augmentation	Training word embeddings	Defining subspace and manipulating word embeddings
Pros	Repeatable by different researchers	Repeatable by different researchers. Easy to replace pronouns.	Allowing personalized settings suitable for target word embeddings	Repeatable by different researchers Easy to manipulate trained word embeddings
Cons	Will lose some information in the removed data and cannot remove indirect bias. Need to retrain the word embeddings after removal.	Will change the distributions of actual data. Better option: Counterfactual Data, Substitution, which changed half of the data and remain the same distribution. Need to train the word embeddings	Different settings and word embeddings will result in different results. Ignoring indirect bias. Need to train the word embeddings	Ignoring indirect bias.

04

PART FOUR

Discussion

Contributions & Future work

Validity

Internal

The reasons of the outcomes and reduce other unanticipated reasons

Content validity:

More attributes other than genders, relations investigation in word embeddings (indirect bias)

Criterion-related validity:

Different measurements comparisons: which are more sensitive to the bias than others. Do they have similar outcomes?

Construct validity:

What are the constructs of social biases?

Neutrality, equality, related mediators, hierarchy, components?

External

Can be applied to other people and other situations.

Tested with other word embeddings and datasets

Q & A

DREXEL UNIVERSITY
College of
Computing &
Informatics

THANKS

REPORTER： Lu Wang

Advisor: Dr. Jina Huh-Yoo
PhD Program in Information Science
College of Computing and Informatics