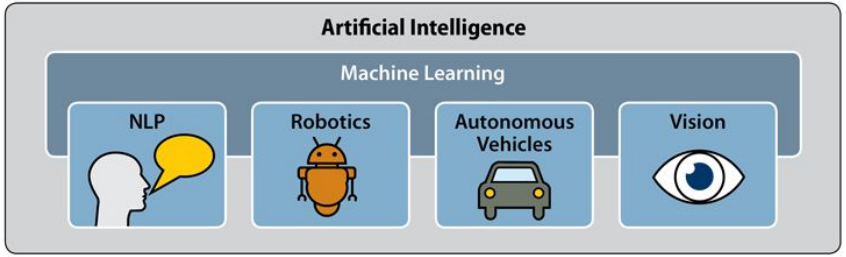
Gender inequality in Italian-language Natural Language Processing datasets

Elia Schneider - Harrisburg University

What is AI social impact?

Artificial Intelligence algorithms have become central in the development of new technologies

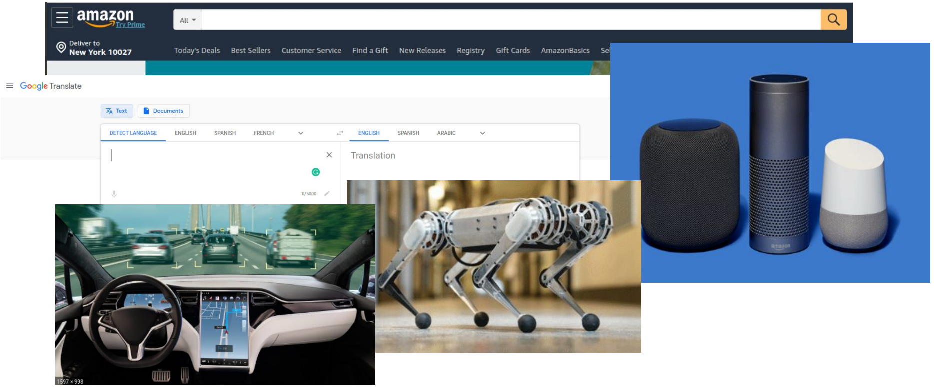


One of the axioms has been:

more data is equal
to better ML algorithm

On the other side
we should not forget that:

with great power
comes great responsibility



https://www.datamation.com/artificial-intelligence/examples-of-artificial-intelligence.html

Bias and fairness in AI

The use of AI raises questions about fairness, transparency, and due process in government decisions and adopted public policies.

PROPUBLICA Graphics & Data Newsletters About

Labor Trump Administration Criminal Justice Health Care More... Ser

MACHINE BIAS

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

REUTERS Business Markets World Politics TV More

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

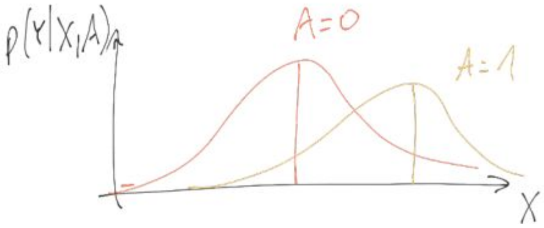
Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

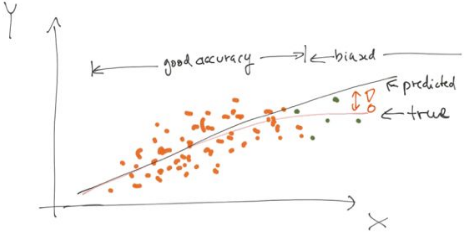
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Why is AI biased?

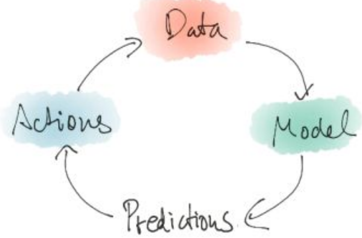
1. **The world and our society are biased**:
biases already present in the datasets;
examples: gender gap, salary and zip code, ...



2. **Underrepresentation in data**:
lack of data exploration during the building of datasets that could leave minority groups underrepresented



3. **Self-fulfilling prophecies**:
biases in how the algorithm minimizes its global error, targeting only the majority groups in the training set



Chouldechova, A., & Roth, A. 2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810
https://cdn.oreillystatic.com/en/assets/1/event/300/Fair%2C%20privacy-preserving%2C%20and%20secure%20ML%20Presentation%201.pdf

Gender discrimination in NLP

25 July 2018

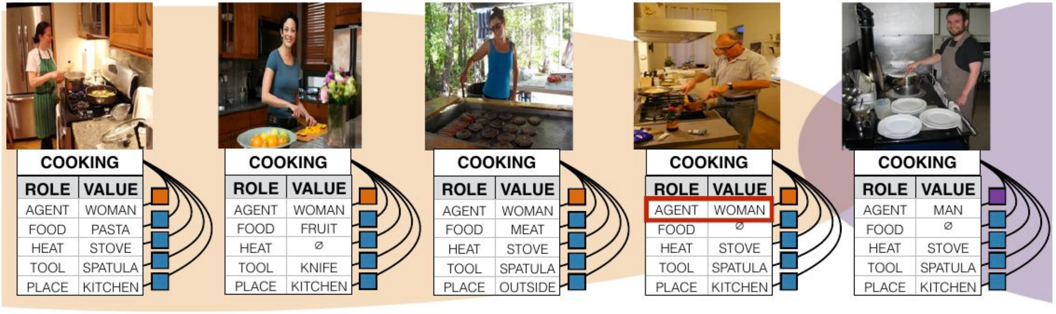
Artificial intelligence is demonstrating gender bias – and it's our fault

Dr Muneera Bano, Lecturer in Software Engineering, Swinburne University of Technology

The data being used to train AI programmes is often gender-biased

www.kcl.ac.uk/news/artificial-intelligence-is-demonstrating-gender-bias-and-its-our-fault

The amplification of gender bias in ML algorithms is more evident when they are trained on datasets based on text, because normally they are huge collections of text produced directly by other humans



Hendricks et all 2018). Women also snowboard: Overcoming bias in captioning models. In European conference on computer vision

Master thesis goals

1. Reproduce methodologies used to measure gender bias in English datasets

2. Adapt already established methodologies to Italian

3. Compare gender bias scores between Italian and English datasets

How to quantify gender bias

A. Word Embedding Association Test (WEAT):

Application of the Implicit Association Test (IAT) used by psychologists to measure subconscious bias in humans, to measure the differences in the strength of association of concepts between genders in NLP datasets.

Caliskan, A., Bryson, J. J., & Narayanan, A. 2017). Semantics derived automatically from language corpora contain human-like biases. Science6334 183

B. Gender direction in an embedding:

Correlation between gender words and neutral words in an embedding matrix.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. 2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 4349–4357.
Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. (NAACL'19).

C. Differences in NLP algorithm performances based on gender

Randomly exchange words related to gender with the other gender in test datasets.

Park, J. H., Shin, J., & Fung, P. 2018). Reducing gender bias in abusive language detection. In pp. 2799–2804). Retrieved from https://aclweb.org/anthology/D18-1302

A. Word Embedding Association

Cosine similarity in word embedding:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}.$$

Word Embedding Association Test score:

$$\mathrm{assoc}(w, A, B) \equiv mean_{a\in A}\Big[\cos(w,a)\Big] - mean_{b\in B}\Big[\cos(w,b)\Big]$$

$$\mathrm{WEAT}(X, Y, A, B) \equiv \sum_{x\in X} \mathrm{assoc}(x, A, B) - \sum_{y\in Y} \mathrm{assoc}(y, A, B)$$

Word Embedding Factual Association Test:

$$\mathrm{WEFAT}(w, A, B) \equiv \frac{mean_{a\in A}\Big[\cos(w,a)\Big] - mean_{b\in B}\Big[\cos(w,b)\Big]}{std_{c\,\in\, A\cup B}\Big[\cos(w,c)\Big]}$$

X & Y gender sets

	Experiment 1
male	john, paul, mike, kevin, steve, greg, jeff, bill, kevin
female	amy, joan, lisa, sarah, diana, kate, ann, donna, amanda
	Experiment 2
male	brother, father, uncle, grandfather, son, he, his, him
female	sister, mother, aunt, grandmother, daughter, she, hers, her
	Experiment 3
male	male, man, boy, brother, he, him, his, son
female	female, woman, girl, sister, she, her, hers, daughter

A & B Attribute sets

	Category 1
work	executive, management, professional, corporation, salary, office
home	home, parents, children, family, cousins, marriage
	Category 2
math	math, algebra, geometry, calculus, equations, computation, numbers
art	poetry, art, dance, literature, novel, symphony, drama
	Category 3
science	science, technology, physics, chemistry, einstein, nasa, experiment
literature	poetry, art, shakespeare, dance, literature, novel, symphony

B. Gender direction in embedding

Cosine similarity in word embedding:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}.$$

Gender direction of a word:
difference between the cosine similarities with two opposed gender-specific words.

Dir(*apples*) = cos(*apples, he*) - cos(*apples, she*)

Gender direction word embedding:
The gender direction average over all neutral gender words.

Gender neighbours for a word (indirect bias):
Ratio male/female gender direction among the first 100 closest words by cosine similarity.

Procedure

1. **Get a clean list of the words in embedding:**
   no numbers, no string too long, ...
2. **Get a list of gender oriented words**
3. **Create a list of words from the embedding that are clean and gender neutral**
4. **Compute bias projections on gender pair words for all gender neutral words in embeddings**
5. **Sum each contribution, cluster most extreme polarized gender neutral words**
6. **Compute gender neighbours for a specific subset of words**

C. Gender performance difference

False positive/negative equality difference:
FPR and FNR are the overall false positive and negative rates and T={male, female}

$$FPED = \sum_{t \in T} |FPR - FPR_t|$$

$$FNED = \sum_{t \in T} |FNR - FNR_t|$$

Hate Speech (HS) detection methodology used:

- Linear Support Vector Machine (SVM)
- Features used:
  - word n-grams in the range 1–3
  - character n-grams in the range 2–4
  - sentence embeddings by average

Procedure

1. **Train SVM classifier using Hate Speech dataset and embedding**
2. **Evaluate classification performances on test dataset**
3. **Generate synthetic test dataset using template:**

   {verb pos/neg} {adjective_pos/neg} {gender}
   being a {gender} is {adjective_pos/neg}
   {name} is a {adjective_pos/neg} {gender}
   you are a {adjective_pos/neg} {gender}
   {pronoun} is a {adjective_pos/neg} {occupation}

4. **Apply trained ML to synthetic test dataset and compute FPED and FNED**

How to apply to Italian

Main differences between English vs. Italian in evaluating gender bias

Gender conjugation of words

English: *Teacher*	Italian: *Maestra/Maestro*

A. Word Embedding Association Test (WEAT):

Most of the words used are gender neutral, only in WEFAT are female and male versions of the same word compared; translate English sets to Italian.

B. Gender direction in an embedding:

Find all pairs female/male words in embedding and use it to compute gender direction:

$$\mathrm{Dir}(maestr^*) = \cos(maestro,\ lui) - \cos(maestra,\ lei)$$

```
{'tore_': 'trice_', 'trice_': 'tore_'}, # attore - attrice
{'tori_': 'trici_', 'trici_': 'tori_'}, # attori - attrici
{'o_':'a_', 'a_':'o_'},                 # bello - bella
{'i_':'e_', 'e_':'i_'},                 # piccoli - piccole
{'e_': 'essa_', 'essa_': 'e_'},         # sacerdote - sacerdotessa
{'i_': 'esse_', 'esse_': 'i_'}          # sacerdoti - sacerdotesse
```

C. Differences in Hate Speech detection:

Create a new template to build synthetic test set

{verb_pos/neg} {det_pron} {gender} {adjective_pos/neg}
essere {undet_pron} {gender} è {adjective_pos/neg}
{name} è {undet_pron} {gender} {adjective_pos/neg}
sei {undet_pron} {gender} {adjective_pos/neg}
{pronoun} è {undet_pron} {occupation} {adjective_pos/neg}

Datasets & Embeddings

English

Embeddings

- Word2vec Google News text dataset

Mikolov T. et all. ICLR, 2013

- Glove embedding based on wiki

Pennington J. et all, 2014

Hate Speech Dataset

- *StormfrontWS*: ~10000 texts from white supremacy forum

https://github.com/aitor-garcia-p/hate-speech-dataset

- *Davidson et. all*: ~25000 tweets tagged whether hate speech, offensive or none

https://github.com/t-davidson/hate-speech-and-offensive-language

Italian

Embeddings

- Italian NLP Lab embedding

http://www.italianlp.it/resources/italian-word-embeddings/

- Word2vec based on wiki

Berardi et all, Word embeddings go to italy

- Specialized embedding for hate speech

Merenda et all,CLIC-IT 2018

Hate Speech Dataset

- *HaSpeeDe*: 4000 Tweets and 4000 Facebook posts collections

http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html

- *HSC*: Italian Twitter Corpus of Hate Speech against immigrants, ~1500 tweets

https://github.com/msang/hate-speech-corpus

WEAT English embeddings

Work vs. Home

	w2v Google news				Glove - wiki			
	Exp 1	Exp 2	Exp 3	Avg	Exp 1	Exp 2	Exp 3	Avg
WEAT score:	0.61	0.76	0.46	**0.61**	1.46	1.04	0.71	**1.07**
Effect size:	1.48	0.67	0.45	**0.86**	1.76	0.93	0.78	**1.16**
P-value:	<0.01	0.09	0.05		<0.01	0.04	0.06	

Math vs. Art

	w2v Google news				Glove - wiki			
	Exp 1	Exp 2	Exp 3	Avg	Exp 1	Exp 2	Exp 3	Avg
WEAT score:	0.16	0.23	0.23	**0.21**	0.59	0.20	0.24	**0.34**
Effect size:	0.68	1.18	0.88	**0.91**	1.50	0.88	0.83	**1.07**
P-value:	0.05	0.01	0.05		<0.01	0.05	0.05	

Science vs. Literature

	w2v Google news				Glove - wiki			
	Exp 1	Exp 2	Exp 3	Avg	Exp 1	Exp 2	Exp 3	Avg
WEAT score:	0.10	0.30	0.25	**0.22**	0.63	0.32	0.32	**0.42**
Effect size:	0.24	1.24	1.00	**0.82**	1.63	1.30	1.22	**1.38**
P-value:	0.38	0.01	0.03		<0.01	0.01	0.01	

X & Y gender sets

	Experiment 1
male	john, paul, mike, kevin, steve, greg, jeff, bill, kevin
female	amy, joan, lisa, sarah, diana, kate, ann, donna, amanda
	Experiment 2
male	brother, father, uncle, grandfather, son, he, his, him
female	sister, mother, aunt, grandmother, daughter, she, hers, her
	Experiment 3
male	male, man, boy, brother, he, him, his, son
female	female, woman, girl, sister, she, her, hers, daughter

A & B Attribute sets

	Category 1
work	executive, management, professional, corporation, salary, office
home	home, parents, children, family, cousins, marriage
	Category 2
math	math, algebra, geometry, calculus, equations, computation, numbers
art	poetry, art, dance, literature, novel, symphony, drama
	Category 3
science	science, technology, physics, chemistry, einstein, nasa, experiment
literature	poetry, art, shakespeare, dance, literature, novel, symphony

WEAT Italian embeddings

Work vs. Home

	NLP Lab				w2v - wiki			
	Exp 1	Exp 2	Exp 3	**Avg**	Exp 1	Exp 2	Exp 3	**Avg**
WEAT score:	0.37	0.53	0.18	**0.36**	0.85	0.25	0.30	**0.47**
Effect size:	0.75	0.97	0.27	**0.66**	1.85	0.62	0.58	**1.01**
P-value:	0.09	0.05	0.34		<0.01	0.16	0.17	

Math vs. Art

	NLP Lab				w2v - wiki			
	Exp 1	Exp 2	Exp 3	**Avg**	Exp 1	Exp 2	Exp 3	Avg
WEAT score:	0.03	0.54	0.41	**0.33**	0.07	0.28	0.29	**0.21**
Effect size:	0.15	1.66	0.97	**0.92**	0.46	1.22	1.48	**1.05**
P-value:	0.40	<0.01	0.04		0.21	0.02	0.01	

Science vs. Literature

	NLP Lab				w2v - wiki			
	Exp 1	Exp 2	Exp 3	Avg	Exp 1	Exp 2	Exp 3	Avg
WEAT score:	-0.26	-0.01	-0.10	**-0.12**	-0.22	0.07	0.06	**-0.03**
Effect size:	-1.35	-0.02	-0.46	**-0.61**	-1.27	0.30	0.80	**-0.06**
P-value:	0.99	0.50	0.76		0.99	0.29	0.10	

X & Y gender sets

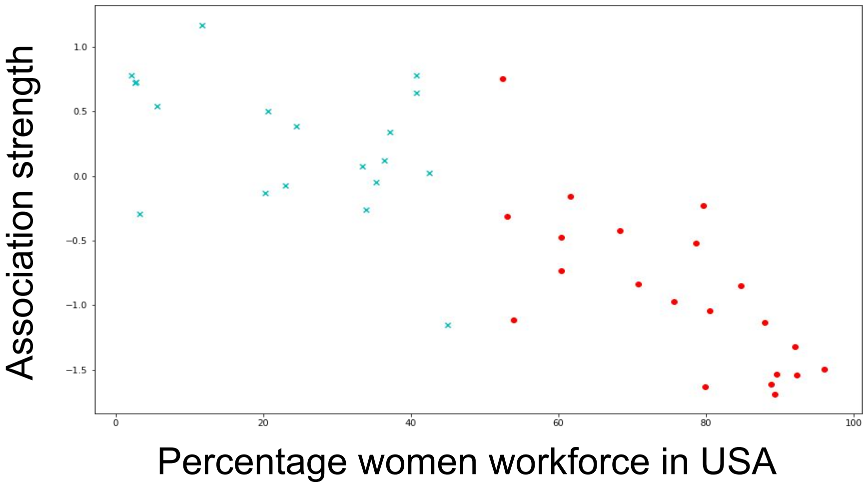
	Experiment 1
male	marco, francesco, alberto, mario, dario, umberto, luigi
female	maria, francesca, caterina, teresa, alice, elena, alessia
	Experiment 2
male	uomo, zio, marito, maschio, padre, nonno
female	donna, zia, moglie, femmina, madre, nonna
	Category 3
male	uomo, suo, ragazzo, fratello, lui, figlio
female	donna, sua, ragazza, sorella, lei, figlia

A & B Attribute sets

	Category 1
work	impresa, professionale, azienda, salario, ufficio, business, carriera
home	casa, genitori, bambini, famiglia, matrimonio, nozze, parenti
	Category 2
math	matematica, algebra, geometria, calculo, equazioni, numeri, addizioni
art	poesia, arte, danza, letteratura, novelle, sinfonia, scultura
	Category 3
science	biologia, fisica, chimica, matematica, geologia, astronomia
literature	filosofia, umanesimo, arte, letteratura, italiano, musica

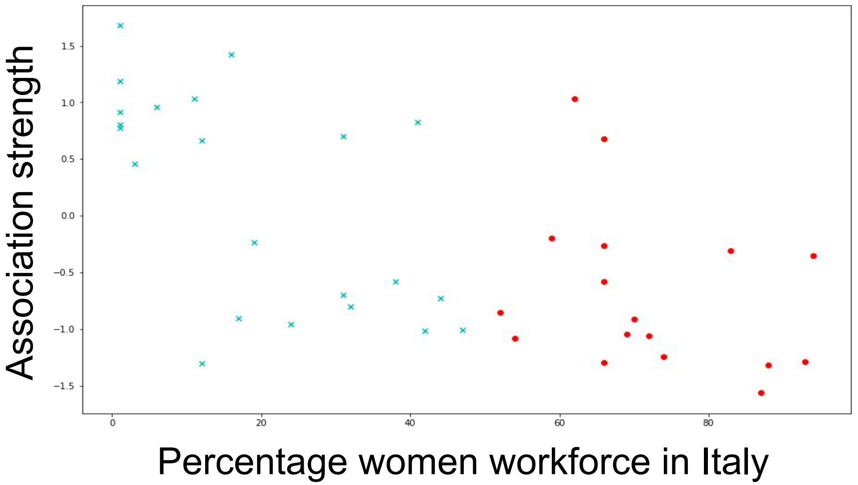
WEFAT gender vs Jobs

English - w2v Google News



Pearson's correlation coefficient: -0.83
P-value < 0.01

Italian - NLP Lab embedding

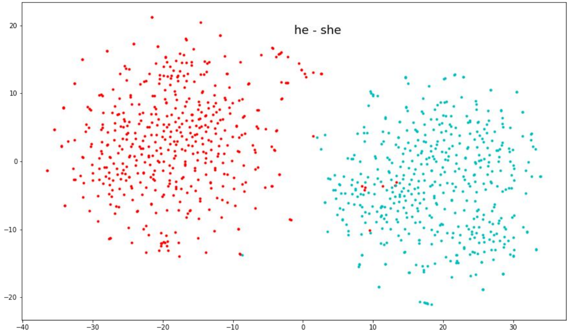


Pearson's correlation coefficient: -0.61
P-value < 0.01

Gender bias in embedding

He - She w2v Google News

overall score: 0.050



Gender cluster precision: 0.99

Top 5 *she* words

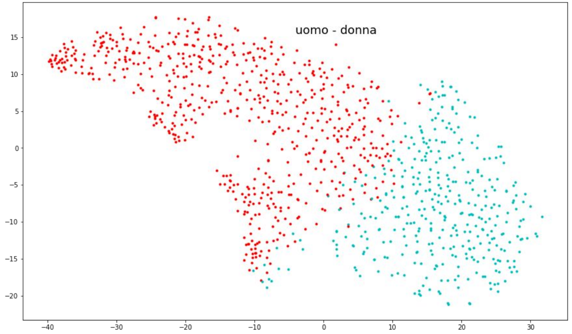
pagaent	-0.377
frontwoman	-0.340
momager	-0.339
pregancy	-0.336
Covergirl	-0.328

Top 5 *he* words

journeyman	0.270
tinkerman,	0.251
outleap	0.250
servicable	0.246
skysports	0.248

Uomo - Donna NLP Lab

overall score: 0.060



Gender cluster precision: 0.74

Top 5 *donna* words

artista	-0.724
angoscia	-0.568
adolescente	-0.555
oste	-0.541
arciere	-0.538

Top 5 *uomo* words

individuo	0.917
intelletto	0.677
destino	0.625
abisso	0.614
personaggio	0.608

Gender bias in HS detection

Dataset	*StormfrontWS*		*Davidson*	*HaSpeeDe*			*HSC* (*)
Embedding	w2v Google news	Glove - wiki	W2v Google news	Specialized Hate	NLP Lab	w2v - wiki	NPL Lab
Tot F-1 score	0.525	0.497	0.273	0.598	0.606	0.639	0.011
Male F-1 score	0.472	0.440	0.306	0.618	0.634	0.663	0.007
Female F-1 score	0.571	0.546	0.236	0.579	0.577	0.616	0.015
FPED	**0.056**	***0.082***	**0.115**	**0.054**	**0.006**	**0.078**	**0.0**
FNED	**0.114**	***0.132***	**0.075**	**0.017**	**0.056**	**0.018**	**0.004**

(*) HSC provides only tweet IDs and I retrieved their text through Twitter's API. Most of the tweets associated with hate are no longer accessible because the user deleted them or Twitter suspended the user. The extremely low F-1 scores are the result of the incomplete datasets where only 14% of the tweets are associated with hate.

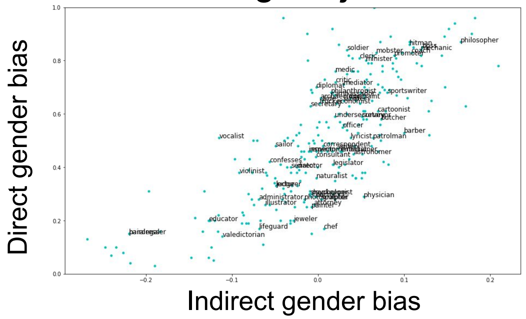
Conclusion and discussion

- **Overall, Italian embeddings have lower gender bias scores, suggesting a minor gender bias in Italian NLP datasets**
  - WEAT test has lower association score and strengths
  - WEFAT shows weaker correlation between gender bias and women percentage workforce
  - Gender direction for Italian slightly higher but lower cluster precision
  - FPED and FNED lower for Italian Hate Speech detection algorithm

- **Issue with applying English gender bias test to Italian**
  - Improvements in gender tagging male/female terms
  - Limitations of existing techniques: expand to indirect bias

- **Gender conjugation reduces direct gender bias in NLP datasets**
  - Association due to conjugation stronger than gender bias
  - Expand analysis to other languages: French. Spanish, …

Bias in English job terms



Thanks for your attention

Additional references:

- Sun, T., et all 2019. Mitigating gender bias in natural language processing: Literature review. 1630–1640. Retrieved from https://www.aclweb.org/anthology/P19-1159
- Collett, C., & Dillon, S. 2019). Ai and gender: Four proposals for future research. Cambridge: The Leverhulme Centre for the Future of Intelligence..
- Barocas, S., Hardt, M., & Narayanan, A. 2019). Fairness and machine learning.fairmlbook.org. http://www.fairmlbook.org
- Bano, M. 2018). Artificial intelligence is demonstrating gender bias - and its our fault. Kings College London News Centre.. Retrieved from www.kcl.ac.uk/news/artificial-intelligence-is-demonstrating-gender-bias-and-its-our-fault
- Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Schwartz, O. 2018). Ai now report 2018. AI Now Institute at New York University

HARRISBURG UNIVERSITY OF SCIENCE AND TECHNOLOGY