Thinking AI – Human Centered AI, October 2021

Footprint of Societal Biases in Natural Language Processing

Navid Rekab-saz

navid.rekabsaz@jku.at @navidrekabsaz

JOHANNES KEPLER UNIVERSITY LINZ

softwarepark hagenberg upper austria

Institute of Computational Perception

Agenda

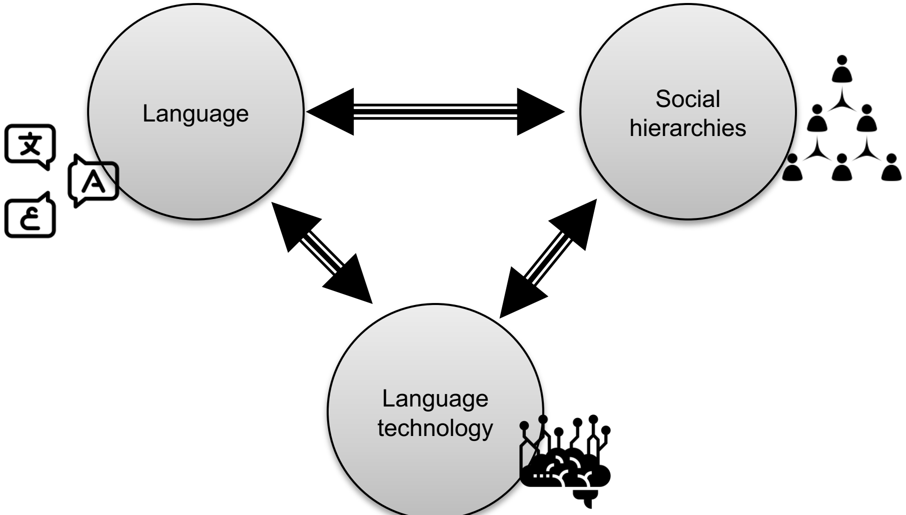
- Bias and Fairness in NLP … what? why?
- Measuring & Monitoring Biases
- Algorithmic Bias Mitigation

Agenda

- **Bias and Fairness in NLP … what? why?**
- Measuring & Monitoring Biases
- Algorithmic Bias Mitigation

Language and Society

- Language …
  - takes on and defines social meaning
  - forms and maintains social hierarchies by …
    - labeling social groups
    - transmitting the beliefs about social groups



Maass, Anne. "Linguistic intergroup bias: Stereotype perpetuation through language." *Advances in experimental social psychology*. 1999.
Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. "Language (Technology) is Power: A Critical Survey of" Bias" in NLP." *In Proc. Of ACL 2020*

Machine Learning (ML) Cycle

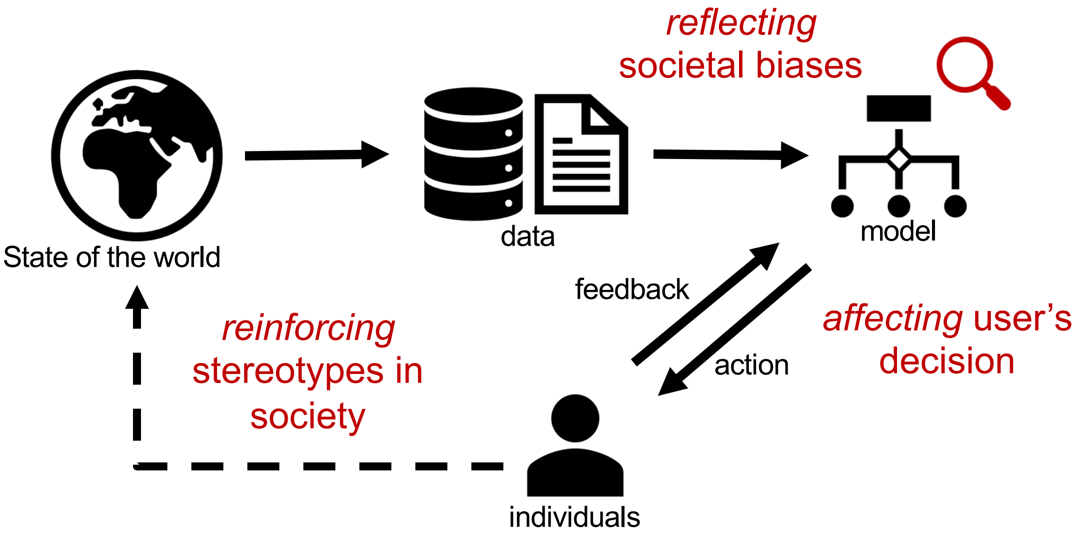
Machine Learning and Societal Biases

ML can observe societal phenomena

- Questions like *“how the perception of girls and boys towards the color pink has changed over time?”*

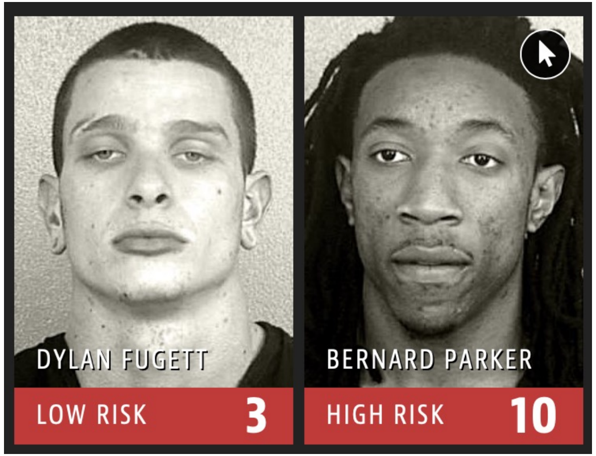
ML can reinforce societal biases

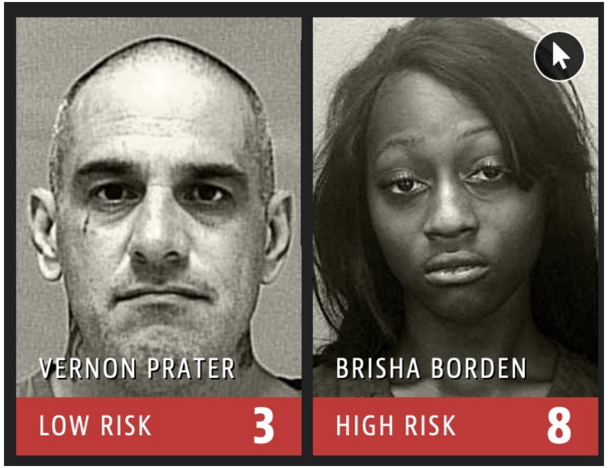
- Encoded societal biases and stereotypes can affect decision making of users and eventually reinforce biases in society



Bias in Crime Discovery

- Predicted risk of reoffending

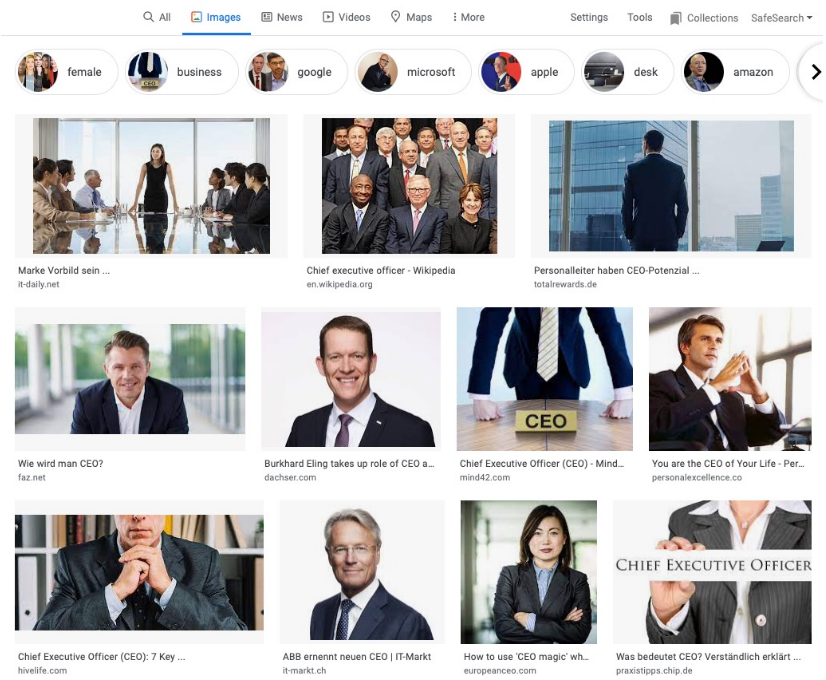




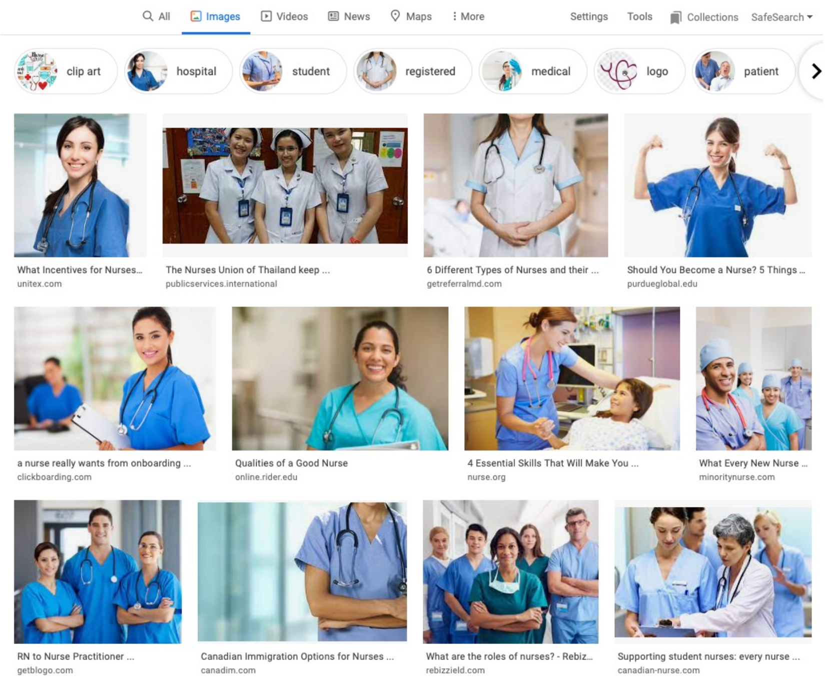
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bias in Search Engines

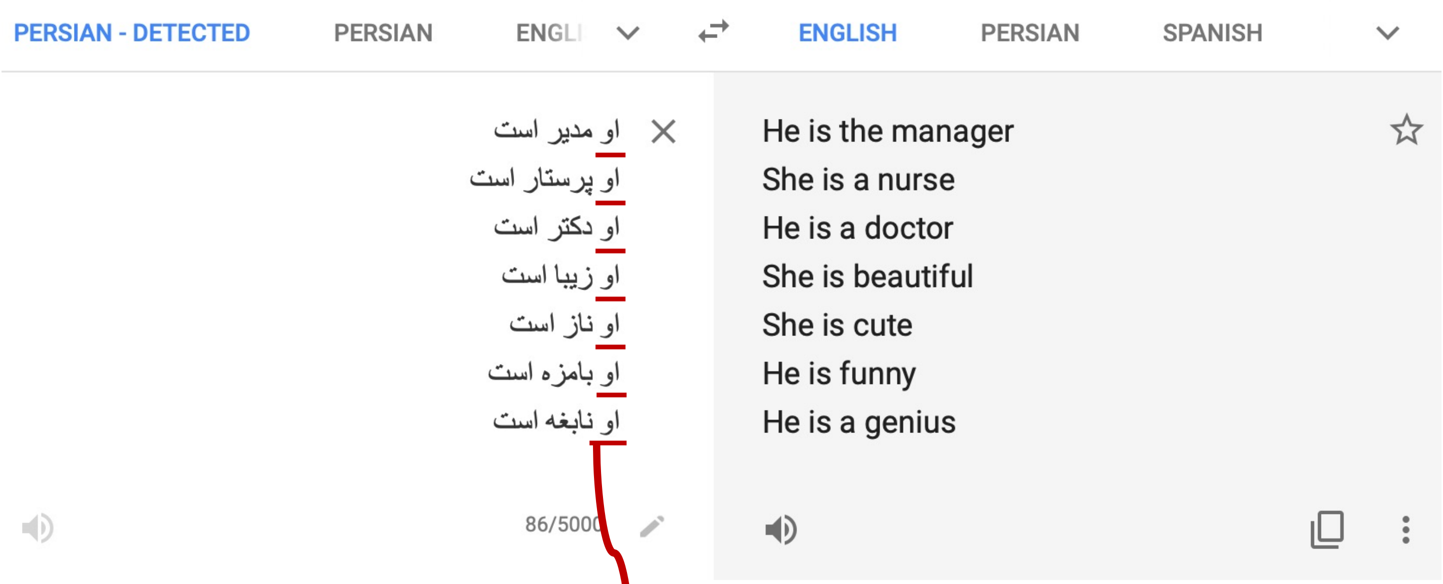
CEO



Nurse



Bias in Automatic Machine Translation



same gender-neutral pronoun

Bias in Image Processing

Google says sorry for racist auto-tag in photo app

https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app

FaceApp's creator apologizes for the app's skin-lightening 'hot' filter

https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology

Beauty.AI's 'robot beauty contest' is back – and this time it promises not to be racist

https://www.wired.co.uk/article/robot-beauty-contest-beauty-ai

Complexity of Studying Bias/Fairness in NLP

A “sample” task – occupation prediction from biographies:

[She/He?] graduated from Lehigh University, with honours in 1998. [Nancy/Adam?] has years of experience in weight loss surgery, patient support, education, and diabetes.

→ Nurse

Language is inherently intertwined with *semantics* and *implicit meanings*

De-Arteaga, Maria, et al. "Bias in bios: A case study of semantic representation bias in a high-stakes setting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019.

What we talk about when we talk about *Bias*

- Biases and stereotypes *per se* do not imply negative connotations.

From “bias”, we mean …

“Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.”

Oxford dictionary

“demographic disparities in algorithmic systems that are objectionable for societal reasons.“

Fairness and Machine Learning

Solon Barocas, Moritz Hardt, Arvind Narayanan, 2019, fairmlbook.org

"I think your test grading is biased in favor of students who answer the test questions correctly."

How *harmful*?!

Allocational harms

- A system allocates resources and opportunities unfairly to different social groups
  - E.g., credit and jobs distribution to minorities

Representational harms

- A system represents some social groups in a less favorable light than others.
  - E.g., stereotyping in a search engine or a recommender system that propagates negative generalizations about particular social groups

Fairness

- What is fair?!
- How is it quantified? Which metrics?
- How can we optimize models for a societal/philosophical concept?!

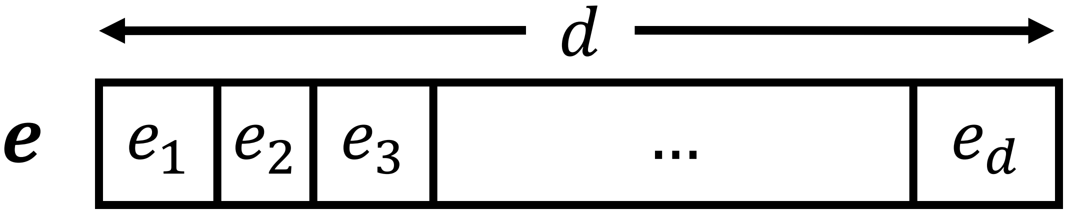
Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. "Language (Technology) is Power: A Critical Survey of" Bias" in NLP." *In Proc. Of ACL 2020*
Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In Proceedings of SIGCIS,

Agenda

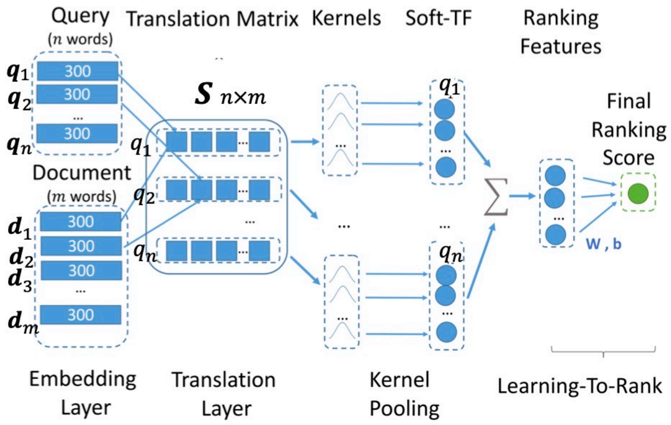
- Bias and Fairness in NLP … what? why?
- **Measuring & Monitoring Biases**
- Algorithmic Bias Mitigation

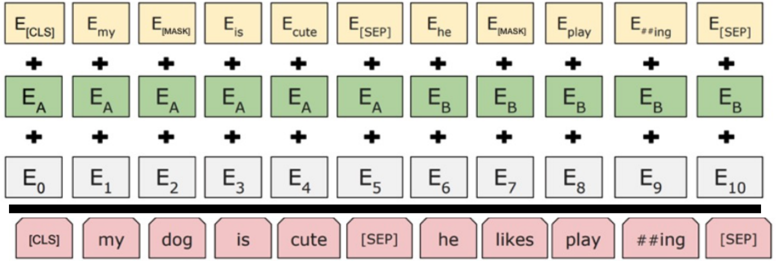
Embeddings!

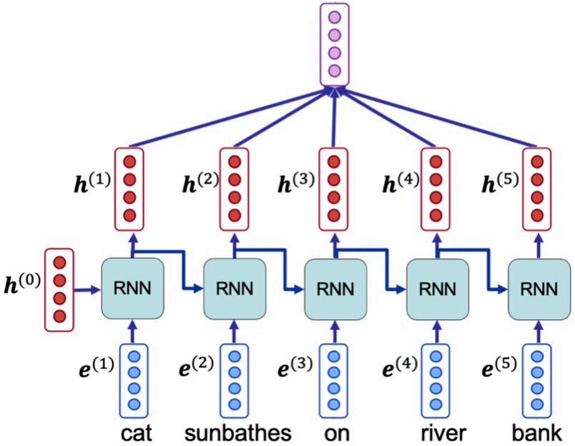
- A word/sentence/document is represented with a vector of d dimensions
- The vector represents the meaning or semantics



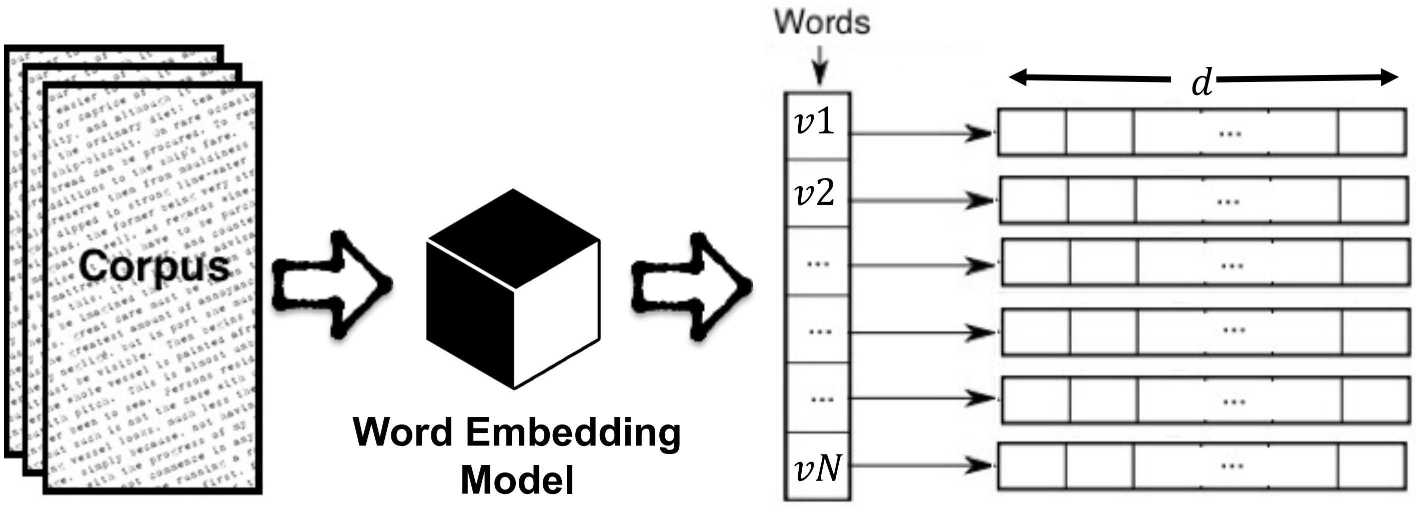
Modern NLP is built on Embeddings

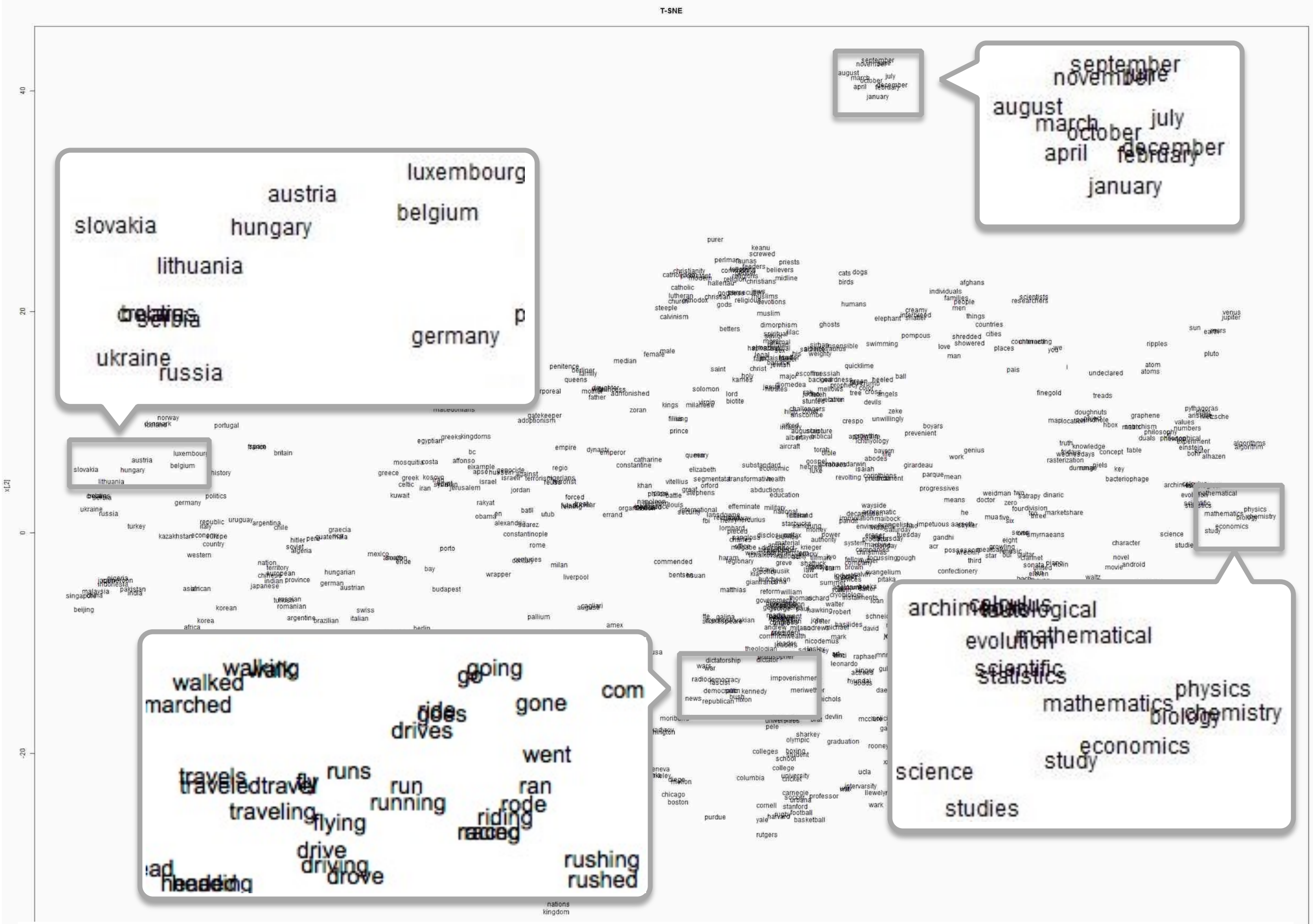






Recipe for Creating Word Embeddings





Word embeddings projected to a two-dimensional space

Semantic Information in Word Embeddings

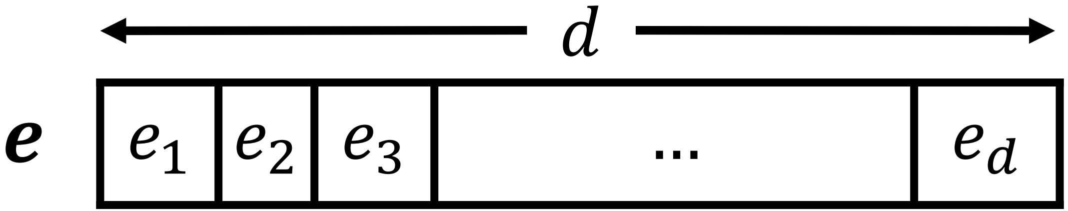
- ***man*** to ***woman*** is like ***king*** to *? (**queen**)*

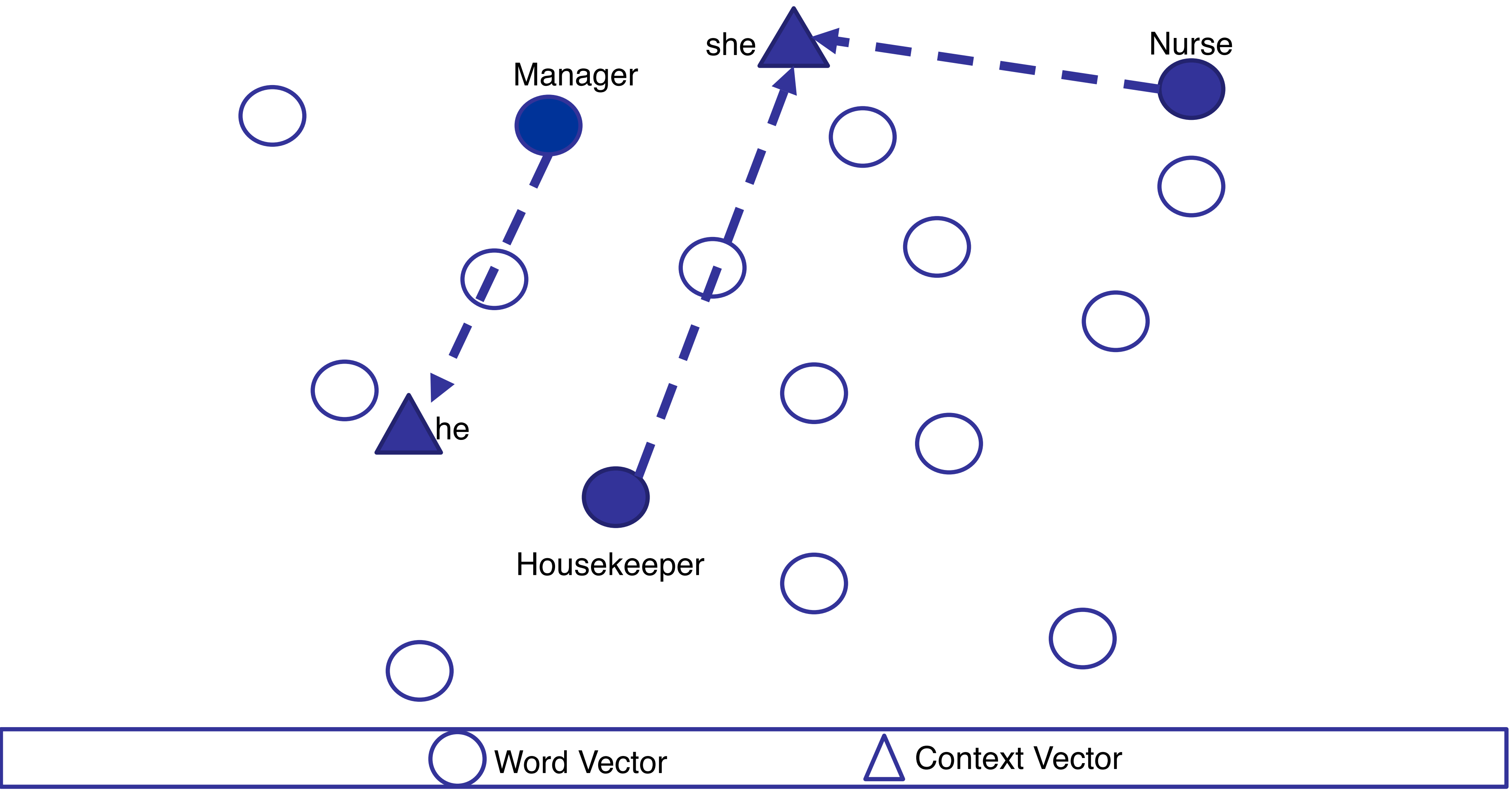
$$\boldsymbol{x}_{\text{king}} - \boldsymbol{x}_{\text{man}} + \boldsymbol{x}_{\text{woman}} = \boldsymbol{x}^*$$
$$\boldsymbol{x}^* \approx \boldsymbol{x}_{\text{que}en}$$

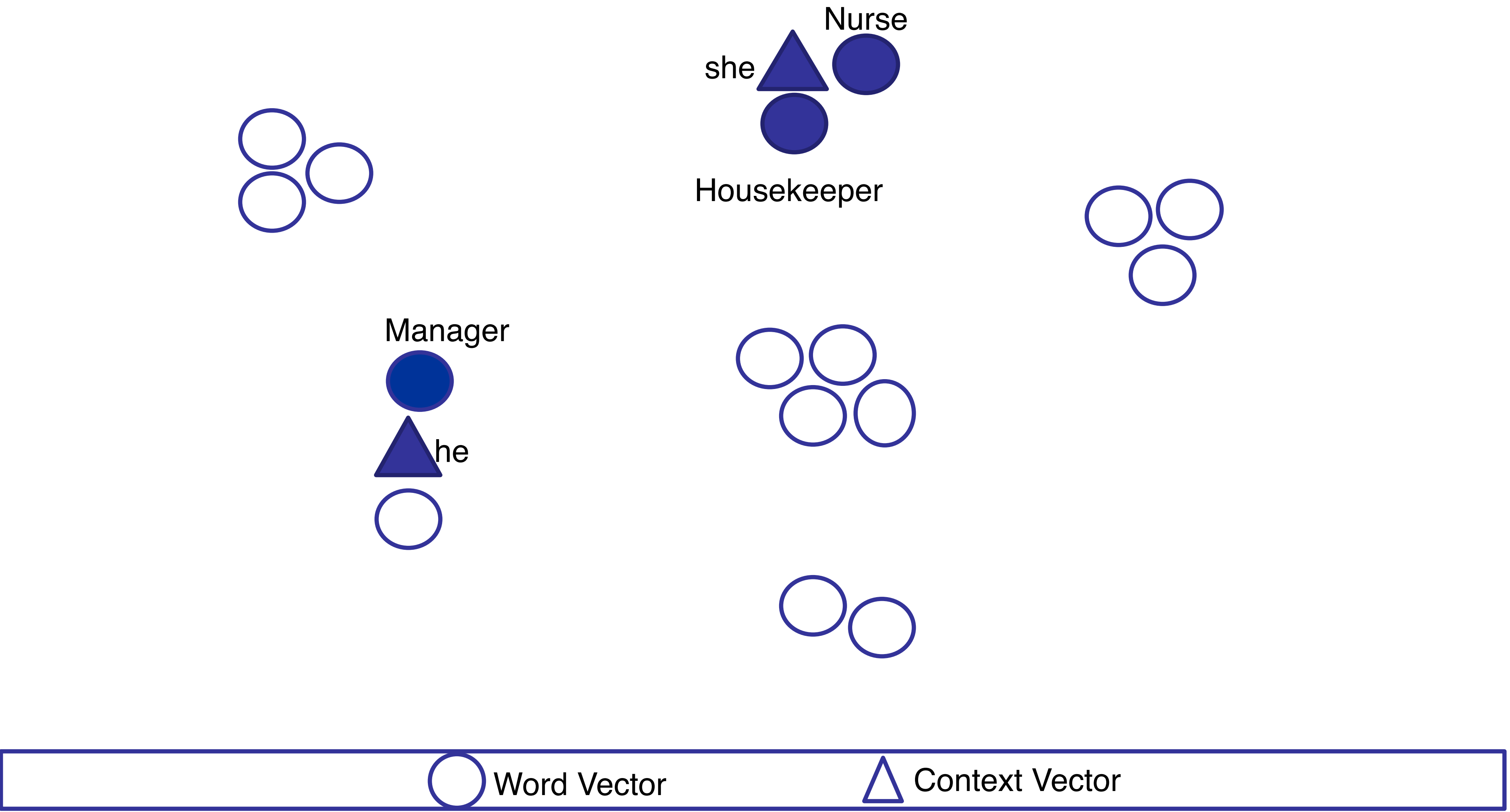
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Embeddings and bias

Representation learning encodes information but also may encode the underlying biases in data!




she
Nurse
Manager
he
Housekeeper
Word Vector
Context Vector


Nurse
she
Housekeeper
Manager
he
Word Vector
Context Vector

Biases reflected in word analogies

- ***she*** to ***he*** is like …

Gender stereotype *she*-*he* analogies

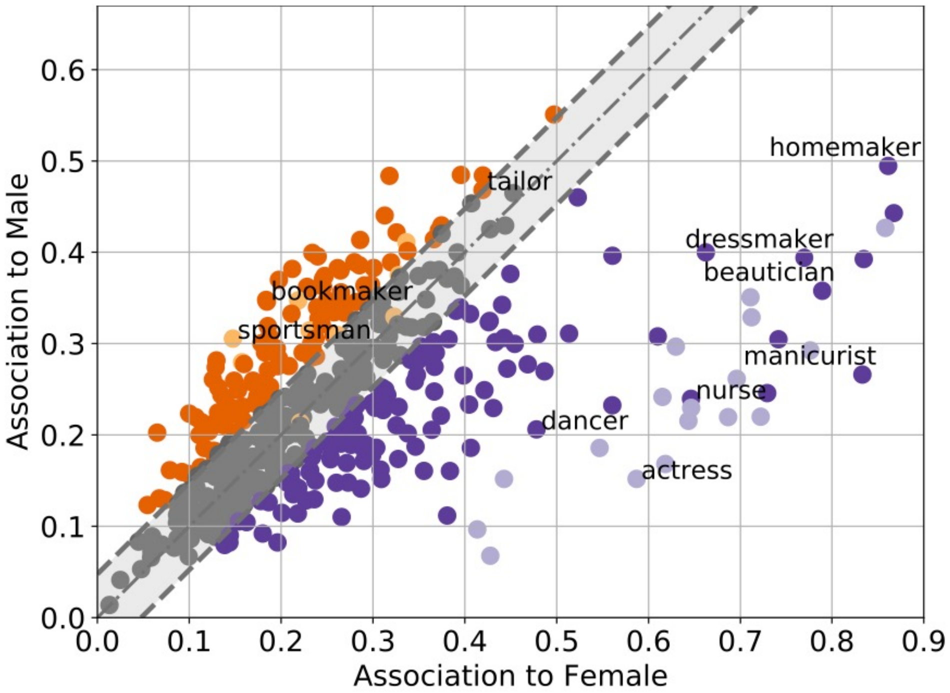
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas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

Gender appropriate *she*-*he* analogies

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

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T.. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (2016)

Biases reflected in Word Embeddings



Associations are measured using a word2vec model, trained on a recent Wikipedia corpus

Navid Rekabsaz, Robert West, James Henderson, Allan Hanbury. Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *In proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2021*

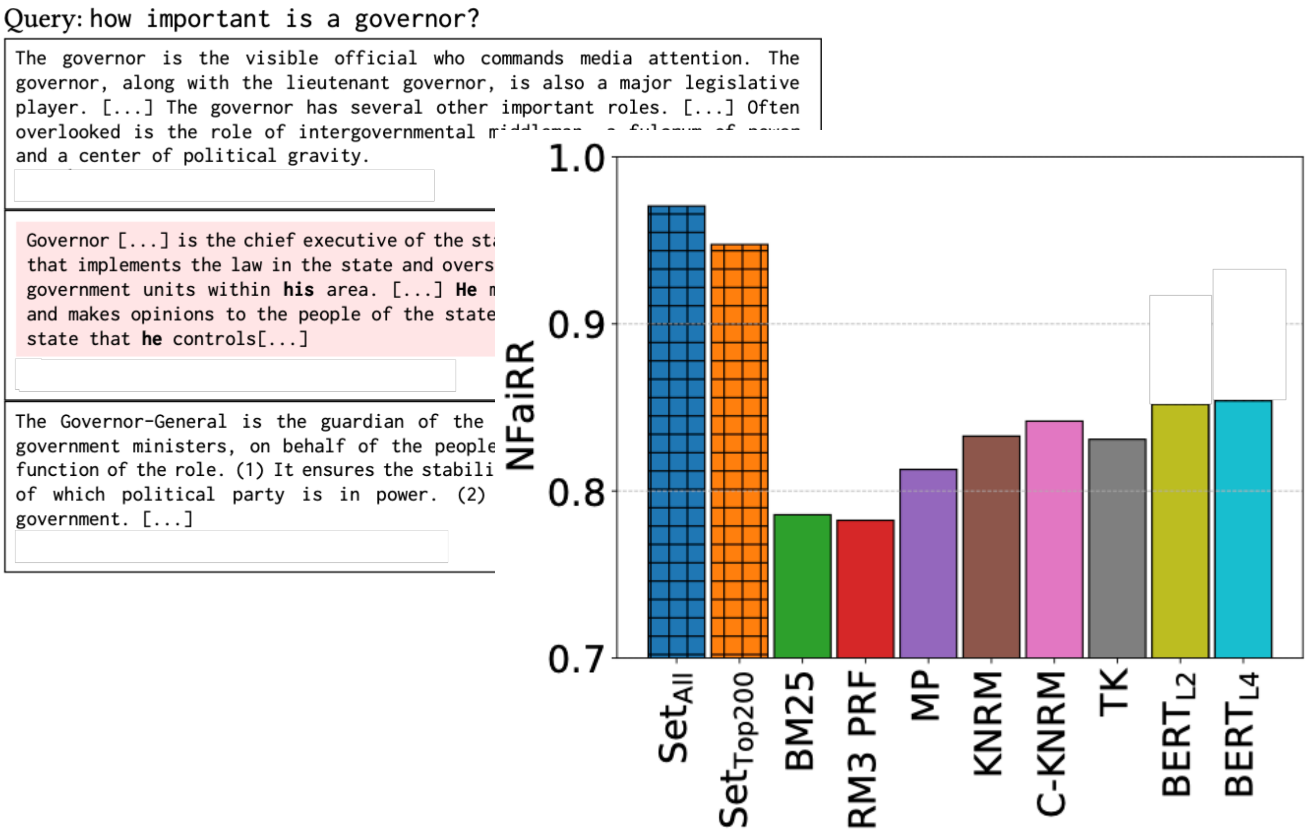
Correlations with job market statistics

Correlation results of the gender bias values, calculated with word embedding to the statistics of the portion of women in occupations

Order	Representation	Method	Labor Data Spearman ρ	Labor Data Pearson's r	Census Data Spearman ρ	Census Data Pearson's r
High-Order	PMI	Directional	0.28	0.07	0.18	0.02
		Centroid	0.14	0.21	0.35	0.40
		Average$_{\text{High}}$	0.33	0.24	0.27	0.19
	PMI-SVD	Directional	0.05	0.07	0.00	0.00
		Centroid	0.41	0.47	0.46	0.53
		Average$_{\text{High}}$	0.41	0.49	0.49	0.56
First-Order	PMI	Average$_{\text{First}}$	**0.53**	**0.51**	**0.57**	**0.62**
High-Order	PPMI	Directional	0.45	0.49	0.39	0.47
		Centroid	0.43	0.46	0.45	0.50
		Average$_{\text{High}}$	0.43	0.46	0.45	0.52
	PPMI-SVD	Directional	0.05	0.07	0.00	0.00
		Centroid	0.41	0.47	0.46	0.53
		Average$_{\text{High}}$	0.41	0.49	0.49	0.56
First-Order	PPMI	Average$_{\text{First}}$	**0.59**	**0.58**	**0.64**	**0.64**
High-Order	SPPMI	Directional	0.26	0.37	0.26	0.28
		Centroid	0.39	0.45	0.45	**0.48**
		Average$_{\text{High}}$	0.32	0.40	0.44	**0.48**
	SPPMI-SVD	Directional	0.17	0.29	0.11	0.03
		Centroid	0.28	0.35	0.39	0.43
		Average$_{\text{High}}$	0.26	0.38	0.36	0.46
First-Order	SPPMI	Average$_{\text{First}}$	**0.57**	**0.49**	**0.52**	**0.48**
High-Order	GloVe	Directional	0.53	0.56	0.34	0.46
		Centroid	0.58	0.60	0.39	0.51
		Average$_{\text{High}}$	**0.60**	**0.60**	0.39	0.51
First-Order	initGlove	Average$_{\text{First}}$	0.38	0.42	0.40	0.51
	eGloVe		0.56	0.57	**0.42**	**0.52**
High-Order	SG	Directional	0.50	0.54	0.58	0.64
		Centroid	0.55	0.57	0.60	0.65
		Average$_{\text{High}}$	0.55	0.57	0.59	0.65
First-Order	eSG	Average$_{\text{First}}$	**0.66**	**0.61**	**0.67**	**0.70**

Fairness in Information Retrieval

a fairness-sensitive query



Navid Rekabsaz, Simone Kopeinik, Markus Schedl. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. *In proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021).*

Agenda

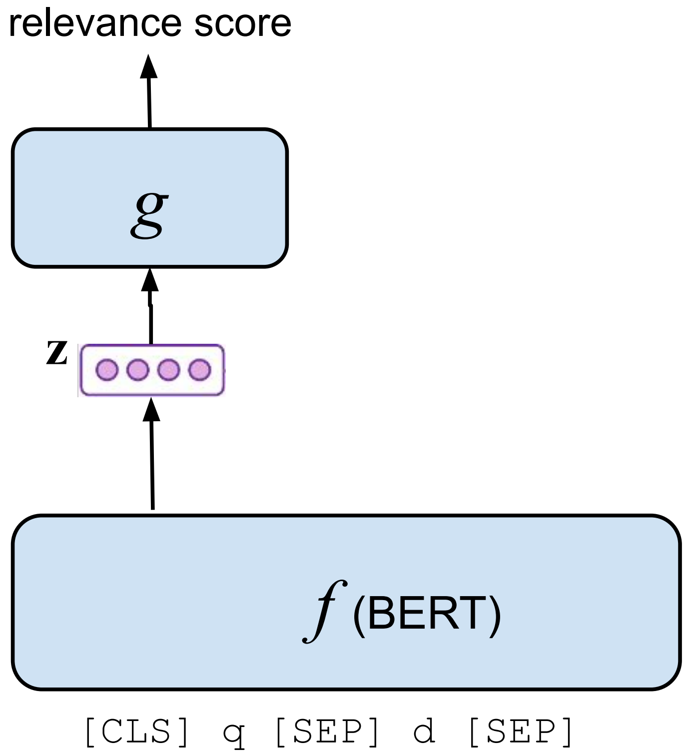
- Bias and Fairness in NLP … what? why?
- Measuring & Monitoring Biases
- **Algorithmic Bias Mitigation**

Algorithmic Bias Mitigation

- Methods to mitigate or reduce bias
  - The aim is to make the output or decision of a model agnostic to sensitive features (such as gender, race, ethnicity, age)

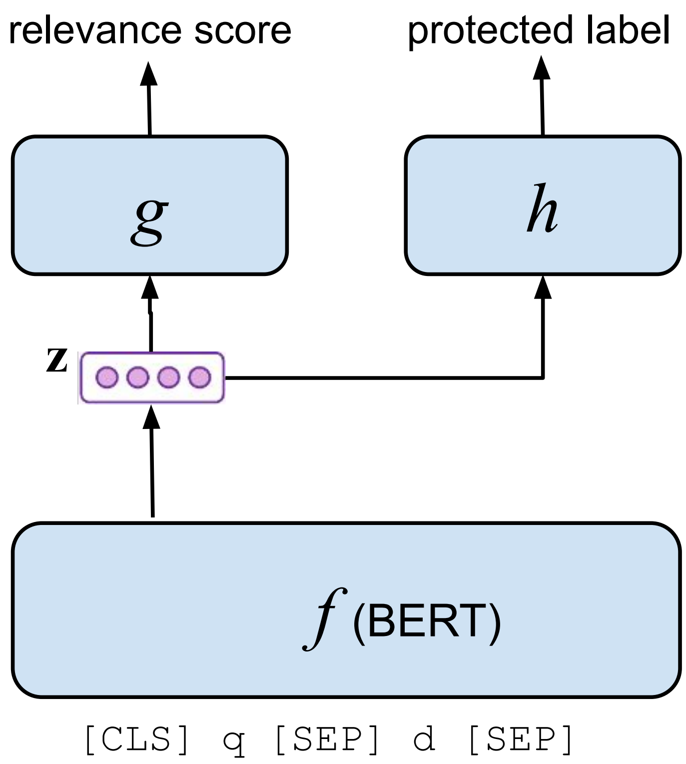
- Categories:
  - Pre-processing: by changing/manipulating dataset
  - In-processing:
    - By adding fairness criteria to model's objective function
    - By training networks that remove sensitive information in learned embeddings
  - Post-processing: by changing/rearranging model's outputs

In-processing Bias Mitigation:
Adversarial Training



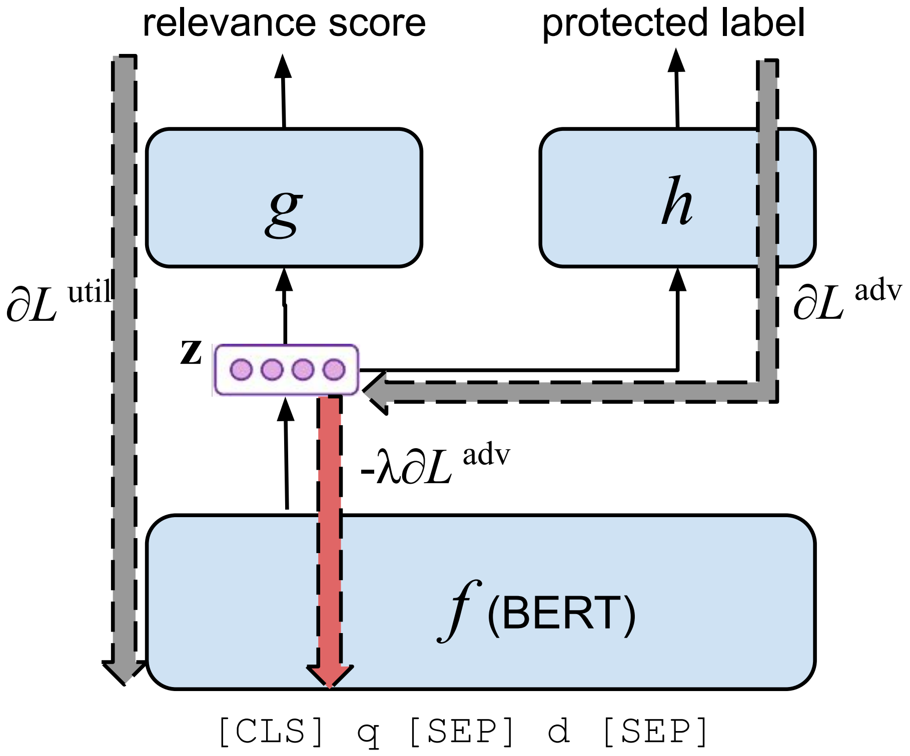
Navid Rekabsaz, Simone Kopeinik, Markus Schedl. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. *In proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021).*

In-processing Bias Mitigation:
Adversarial Training



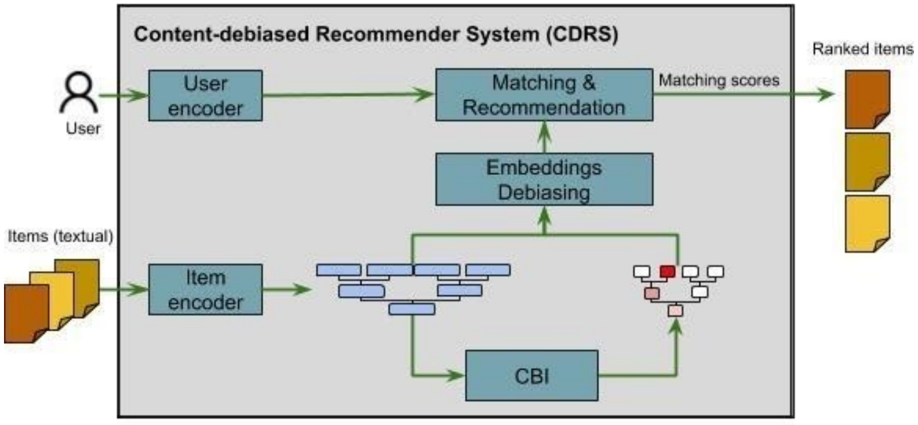
Navid Rekabsaz, Simone Kopeinik, Markus Schedl. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. *In proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021).*

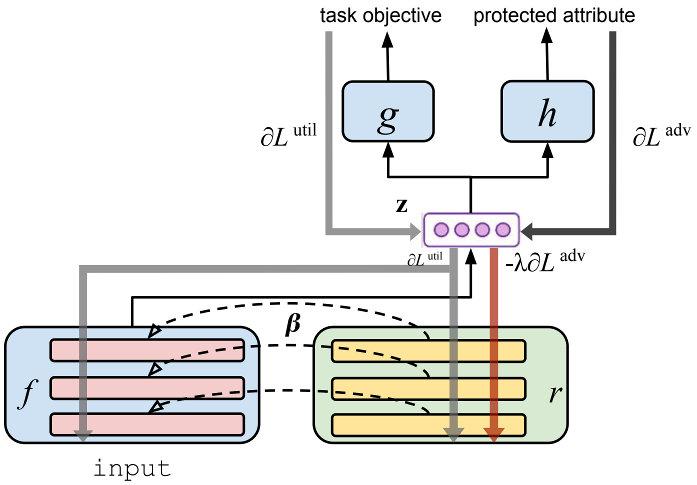
In-processing Bias Mitigation:
Adversarial Training



Navid Rekabsaz, Simone Kopeinik, Markus Schedl. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. *In proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021).*

Fairness through Filtering Bias Flow





- Mitigating Gender Bias in Job Recommender Systems: A Machine Learning-Law Synergy (**TIMELY**)
- Funded by Linz Institute of Technology (**LIT**)

Final words…

- Fairness and bias are social concepts and inherently normative

- Bias in NLP systems should be grounded in its social context

"… without this grounding, researchers and practitioners risk measuring and mitigating only what is convenient to measure and mitigate, rather than what is most normatively concerning."

Blodgett et al. [2020]

- Real problems need interdisciplinary thinking!
  - Addressing bias requires going beyond CS and getting engaged with disciplines such as sociolinguistics, linguistic anthropology, sociology, law, psychology, etc..

Questions?

JKU
JOHANNES KEPLER
UNIVERSITY LINZ

Institute of
Computational
Perception