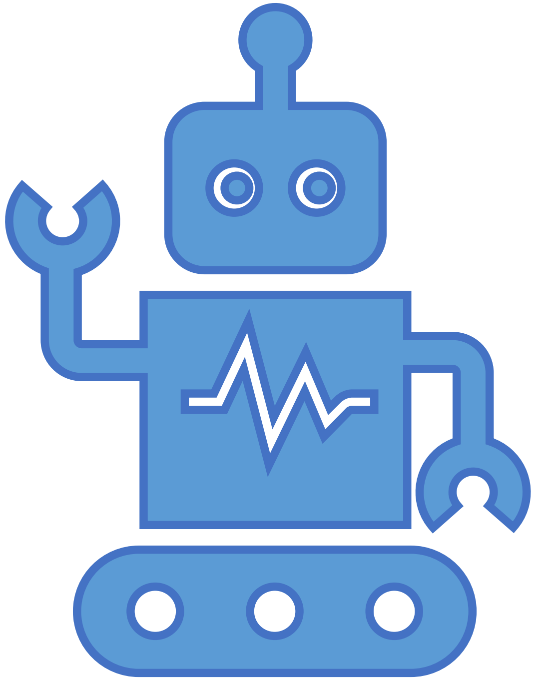
Bias In NLP

Fatma Elsafoury

Agenda

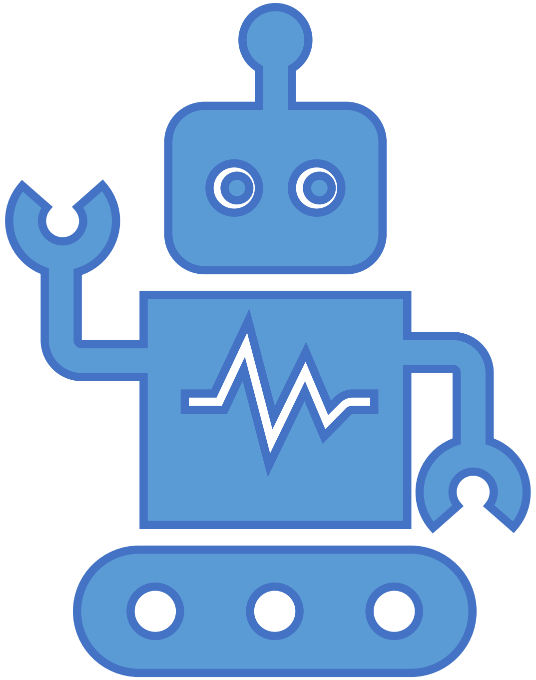


Bias in NLP

#women_in_nlp

talk series

Today's talk



Bias in NLP

#women_in_nlp

talk series

Bias in NLP

- In 2021 Claudia Wagner and co-authors "algorithmically infused societies as the societies that are shaped by algorithmic and human behaviour" like social media platforms [1].
- The data collected from these societies carry the same bias in algorithms and humans, like population bias and behavioural bias [2].
- unsupervised models like word embeddings encode these biases during training [3]

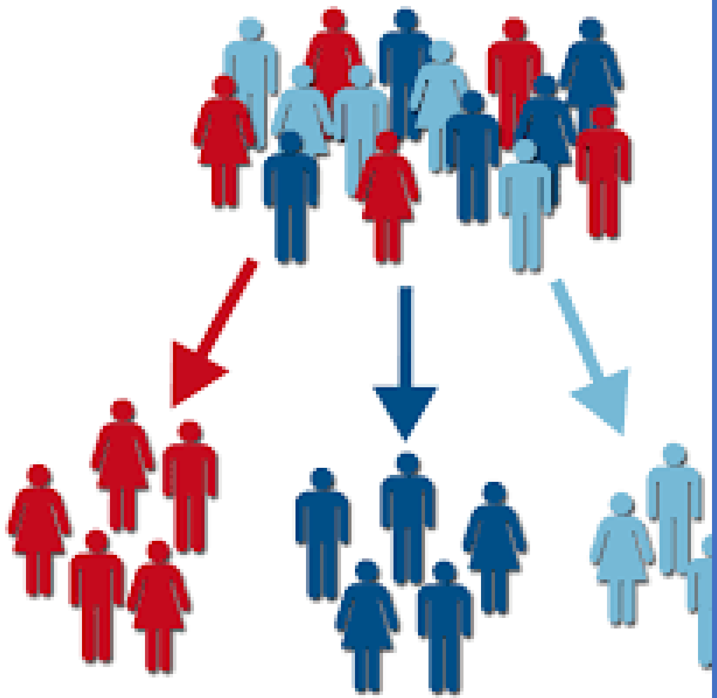
[1] Measuring algorithmically infused societies.

[2] Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

[3] Understanding the Origins of Bias in Word Embeddings

Social Bias

- To group people in predefined categories to make it easier for our brains to deal with them. E.g. Gender and racial bias [4].
- Most studied in the literature of bias in NLP.
- Metrics to measure social bias in word embeddings are WEAT, RNSB, RND, and ECT.



[4] The End of Bias, Nordell 2021

Offensive stereotyping

- Using slurs and swear words to describe groups of people aiming at stressing on the inferiority of the identity of the marginalized group [5].
- The internet is rife with slurs and profanity, it is important to study how machine learning models encode this offensive stereotyping.

[5] Slurs, interpellation, and ideology. The Southern Journal of Philosophy, 56:7–32

Systematic Offensive Stereotyping (SOS) bias

- Statistical definition:
  - A systematic association in the word embeddings between profanity and marginalised groups of people e.g. women, LGBTQ, and non-white-ethnicities.
- We look the SOS bias in 5 word embeddings:
  - Word2vec, glove-wk, glove-twitter, UD, and chan.

Measure SOS bias

- We used Non-offensive Identity words (NOI) to describe different groups of people.
- We used a list of 427 swear words [6].

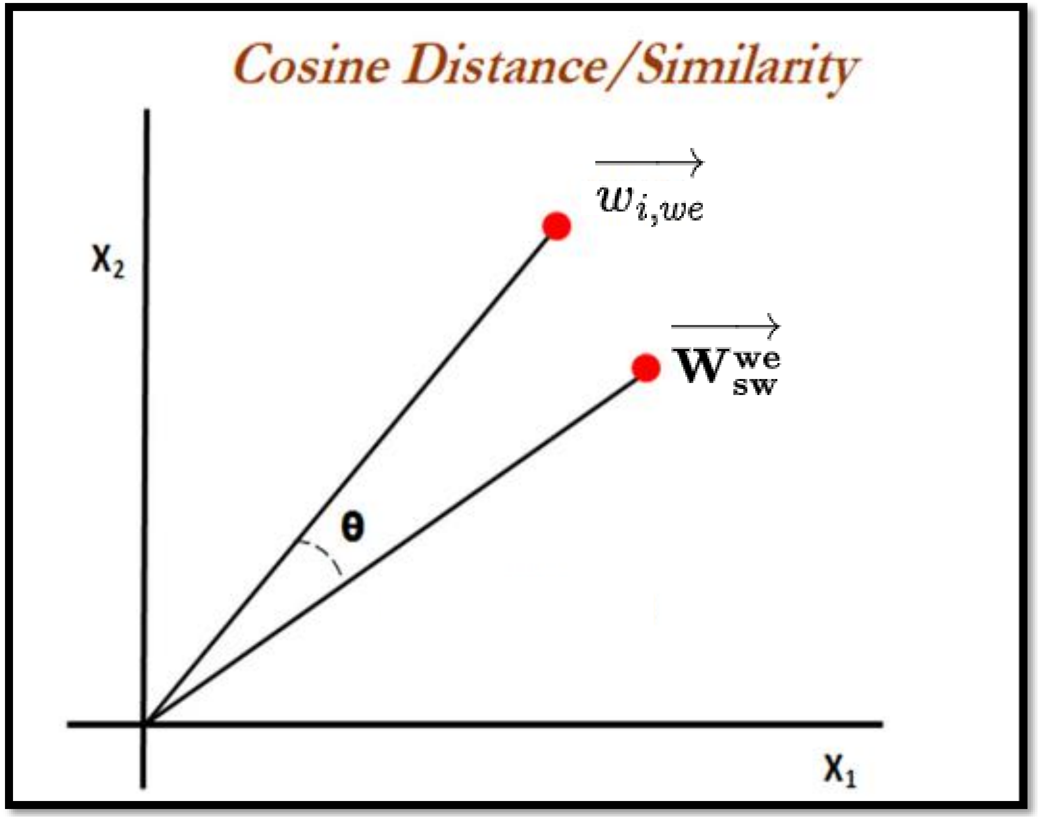
Group	Word
LGBTQ*	lesbian, gay, queer, homosexual, lgbt, lqbtq, bisexual, transgender, tran, non-binary
Women*	woman, female, girl, wife, sister, mother, daughter
Other ethnicities*	african, african american, black, asian, hispanic, latin, mexican, indian, arab,middle eastern
Straight	hetrosexual, cisgender
Men	man, male, boy, son, father, husband, brother
White ethnicities	white, caucasian, european american, european, norwegian, canadian, german, australian, english, french, american, swedish, dutch

*Marginalised group

Table 1: NOI words and the group they describe.

[6] Deep learning for detecting cyberbullying across multiple social media platforms.

Measure SOS bias



we Is a word embeddings model e.g. word2vc, glove-wk, glove-twitter, ud, and chan.

$\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}$ Profanity vector is the average vector of the 427 swear words for a word embeddings we

$\overrightarrow{w_{i,we}}$ Word vector of NOI word for the word embeddings we

$$SOS_{i,we} = cos(\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}, \overrightarrow{w_{i,we}}) = \frac{\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}} \cdot \overrightarrow{w_{i,we}}}{||\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}|| \cdot ||\overrightarrow{w_{i,we}}||}$$

Measure SOS bias

Word embedding	Mean SOS	
	Marginalised	Non-marginalised
Word2Vec	0.403	**0.430**
Glove-WK	**0.448**	0.281
Glove-Twitter	**0.558**	0.461
UD	**0.407**	0.320
Chan	**0.558**	0.393

Table 2: Mean SOS score of the different groups.

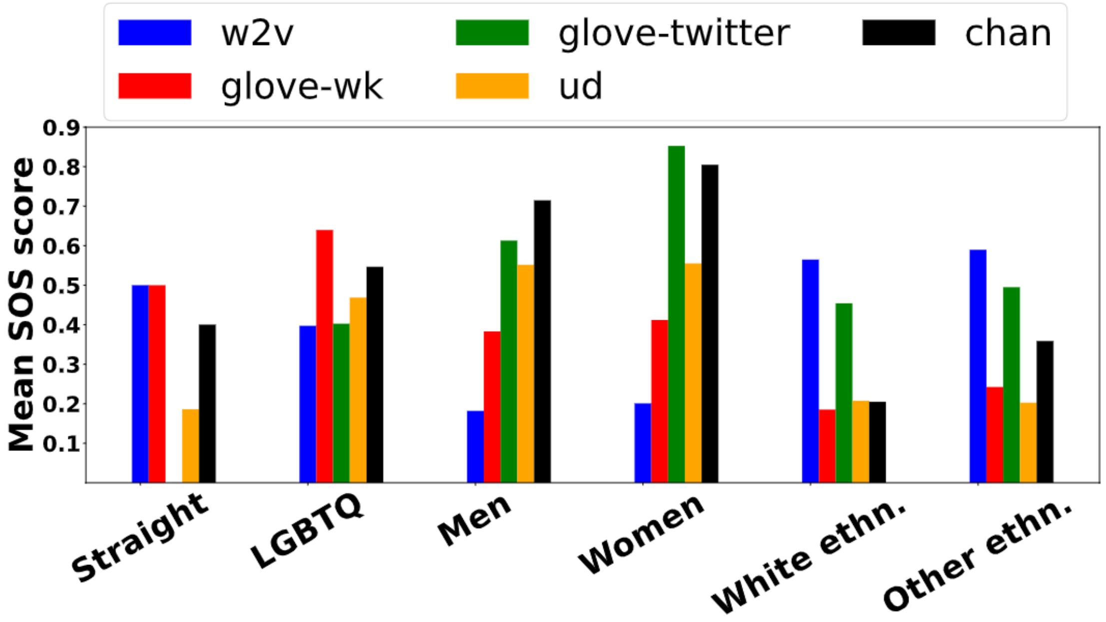


Figure 1: Mean SOS scores for the examined word embeddings and groups.

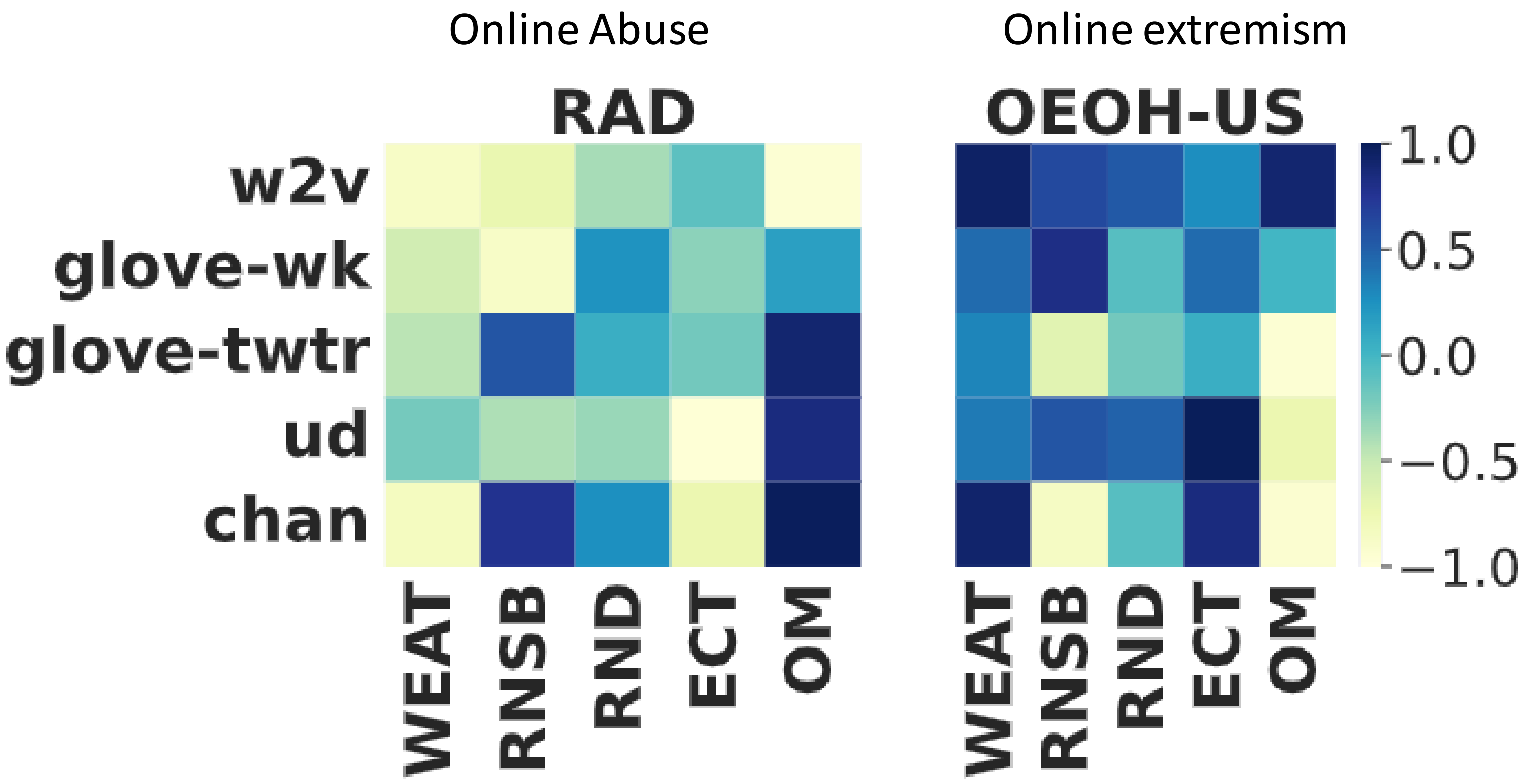
Validating SOS bias

- Compare the SOS bias against published statistics on online abuse and hate against marginalized groups (Women, LGBTQ, and Non-white ethnicities).
  - The RAD Campaign survey on online abuse [7].
  - The survey OEOH online extremism and hate [8].

- Compare our proposed metric to measure the SOS bias against state-of-the-art metrics to measure bias in the literature.
  - WEAT, RNSB, RND, and ECT [9].

[7] Rad Campaign. 2014. The rise of online harassment.
[8] Online extremism and online hate.
[9] WEFE: the word embeddings fairness evaluation framework.

Validating SOS bias



SOS Bias and downstream tasks

- What is the impact of the SOS bias in word embeddings on the downstream task of hate speech detection?
  1. Model performance.
  2. Model unfairness.

SOS Bias and downstream tasks

- Hate speech detection task

Dataset	Samples	Positive samples	Avg. words per comment	Max. words per comment
HateEval	12722	42%	21.75	93
Twitter-sexism	14742	23%	15.04	41
Twitter-racism	13349	15%	15.05	41
Twitter-hate	5569	25%	14.60	32

Note: Positive samples refer to offensive comments

Table 4: Hate speech datasets' details.

SOS Bias and downstream tasks

- Hate speech detection task

Dataset	Model	F1-score				
		Word2Vec	Glove-WK	Glove-Twitter	UD	Chan
HateEval	MLP	0.593	0.583	0.623	0.597	**0.627**
	BiLSTM	0.663	0.651	**0.671**	0.661	0.661
Twitter-sexism	MLP	0.587	0.587	**0.589**	0.578	0.563
	BiLSTM	0.659	0.661	**0.661**	0.625	0.631
Twitter-racism	MLP	**0.683**	0.681	0.680	0.679	0.650
	BiLSTM	0.717	**0.727**	0.6999	0.698	0.712
Twitter-hate	MLP	0.681	0.713	0.775	**0.780**	0.692
	BiLSTM	0.772	0.821	**0.851**	0.837	0.84

Note: Numbers in bold indicate best performance per model and dataset

Table 5: F1 scores for the used models using the examined word embeddings on our datasets.

SOS Bias and model performance

Dataset	Model	Pearson's correlation				
		WEAT	RNSB	RND	ECT	Our_metric
HateEval	MLP	0.84	0.48	0.57	-0.22	**0.88**
	BiLSTM	0.19	-0.10	-0.17	-0.10	**0.42**
Twitter-sexism	MLP	-0.81	-0.99	-0.85	-0.40	**-0.36**
	BiLSTM	-0.44	-0.80	-0.40	-0.61	**0.01**
Twitter-racism	MLP	-0.94	-0.92	-0.96	**-0.12**	-0.62
	BiLSTM	-0.17	-0.08	0.20	**-0.096**	-0.23
Twitter-hate	MLP	-0.13	-0.29	-0.45	-0.25	**0.07**
	BiLSTM	0.57	0.25	0.33	-0.48	**0.67**

Table 12: Pearson correlation coefficient of the SOS bias scores of the different word embeddings and the F1 scores of the used models for each bias metric and dataset.

SOS Bias and downstream tasks

- What is the impact of the SOS bias in word embeddings on the downstream task of hate speech detection?

1. <u>Model performance:</u>
   - Our SOS bias metric is more positively correlated to the model performance than state of the art bias metrics.
   - Results suggest that the bias in word embeddings, especially SOS bias, might lead to better performance on hate speech detection task.

2. Model unfairness.

SOS Bias and model unfairness

- What is model unfairness in our case?
  - For hate speech detection models, unfairness is falsely assign hateful labels to a sentence because the sentence includes terms describing a marginalized group.
- Measure unfairness:
  - Fairness gap = FPR(marginalized) - FPR(non-marginalised).

SOS Bias and model unfairness

- Measure fairness gender gap:
  - Filter out sentences that contain NOI (women) and sentences that contain NOI (men).
  - FPR (women) - FPR (men)

- Measure fairness racial gap:
  - Filter out sentences that contain NOI (ethn) and sentences that contain NOI (white).
  - FPR (ethn) - FPR (white)

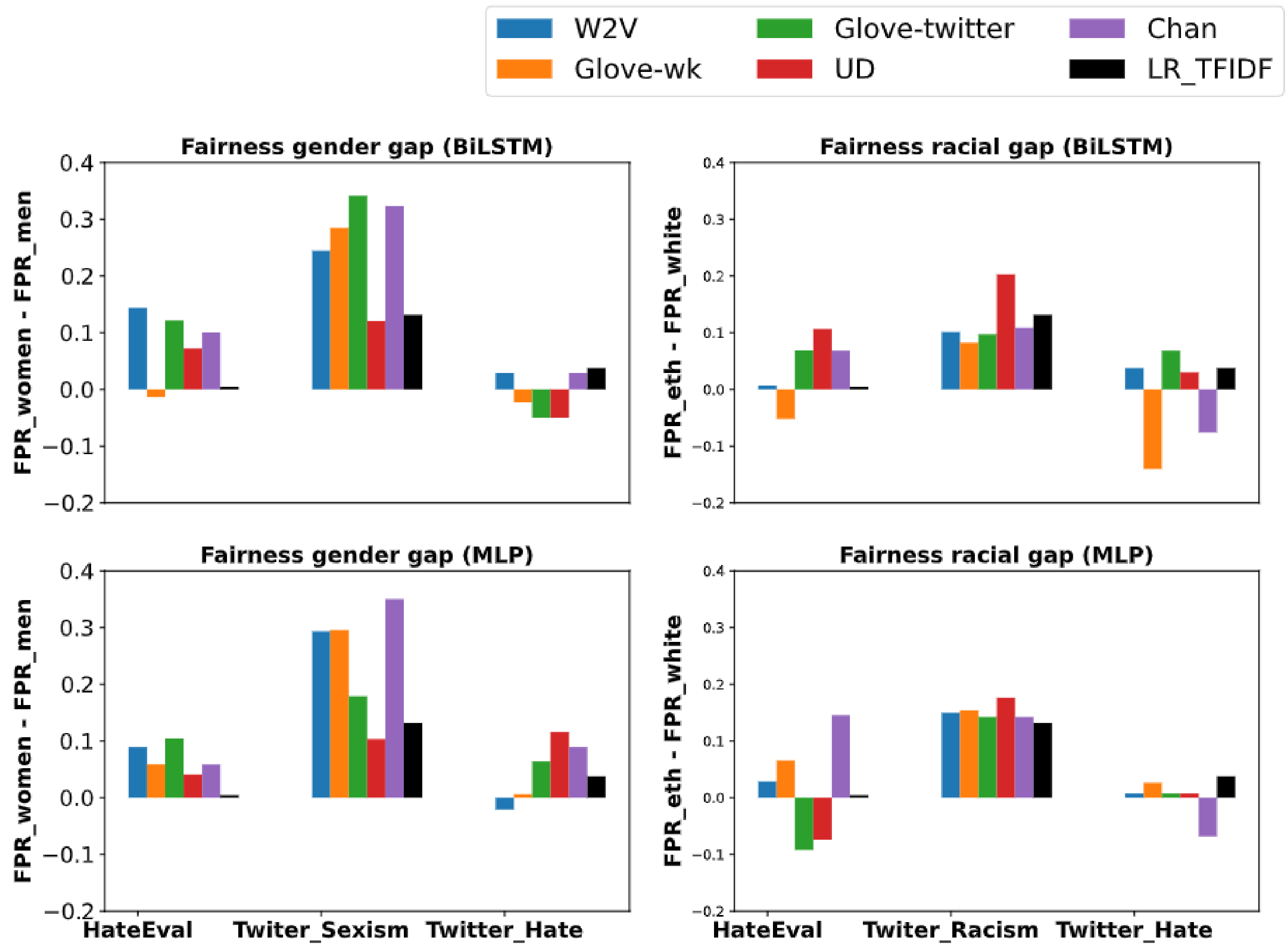


Figure 3: Unfairness scores of the different models and datasets

SOS Bias and model unfairness

- Correlation between bias scores and unfairness scores
  - Gender bias vs. SOS bias
    - Measure Gender bias using WEAT, RNSB, RND, and ECT.
    - Measure SOS (women)
  - Racial bias vs. SOS bias
    - Measure Racial bias using WEAT, RNSB, RND, and ECT.
    - Measure SOS (ethnicity)

SOS Bias and model unfairness (Gender bias)

Dataset	Model	Pearson's correlation				
		WEAT	RNSB	RND	ECT	SOS
HateEval	MLP	-0.002	-0.135	-0.238	-0.745	0.077
	BiLSTM	-0.355	-0.379	0.087	-0.035	0.120
Twitter-sexism	MLP	0.865	0.729	0.629	-0.222	-0.161
	BiLSTM	0.360	0.688	0.432	-0.881	0.436
Twitter-hate	MLP	-0.759	0.028	0.335	0.267	0.728
	BiLSTM	0.666	0.366	0.586	0.155	-0.321

Table 10: Pearson correlation coefficient of the gender bias scores of the different word embeddings and the unfairness gender gaps of the used models for each bias metric and dataset.

SOS Bias and model unfairness (Racial bias)

Dataset	Model	Pearson's correlation				
		WEAT	RNSB	RND	ECT	SOS
HateEval	MLP	0.442	0.664	0.747	-0.192	-0.054
	BiLSTM	-0.750	0.336	0.239	0.533	-0.085
Twitter-racism	MLP	-0.524	-0.338	-0.416	0.712	-0.643
	BiLSTM	-0.790	-0.018	-0.117	0.835	-0.500
Twitter-hate	MLP	0.109	-0.960	-0.967	-0.085	-0.046
	BiLSTM	-0.739	-0.380	-0.408	0.408	0.536

Table 11: Pearson correlation coefficient of the racial bias scores of the different word embeddings and the unfairness racial gaps of the used models for each bias metric and dataset.

SOS Bias and downstream tasks

- What is the impact of the SOS bias in word embeddings on the downstream task of hate speech detection?

1. <u>Model performance:</u>
   - SOS bias is more positively correlated to the model performance than state of the art bias metrics.
   - Results show that the bias in word embeddings , especially SOS bias, might lead to better performance on hate speech detection task.

2. <u>Model unfairness:</u>
   - To some extent the SOS does influence model unfairness especially for gender bias but it is not the case when it comes to racial bias.
   - Other factors contribute models' unfairness like bias in the datasets.
   - Open question and more investigation is needed.

Bias in NLP

Questions?

Today's talk

Bias in NLP

#women_in_nlp

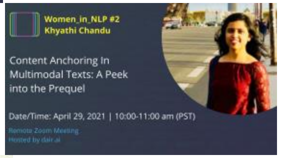
talk series

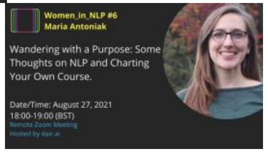
#Women_in_NLP

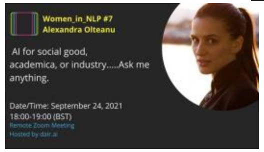
- Shows the findings of some experiements on the effects of being a minority like women in STEM or black people in academia.
  - The cognitive effect leads to a self-fulfilling prophacy.
  - The Physical effect leads to high blood pressure and other complications.
  - To mitigate the negative effects, people need to see representatives of their own group.

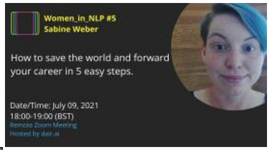
#Women_in_NLP

- Supported by Dair.AI
- Monthly talks on Zoom.
- Events are announced on Meetup and Twitter.
- 10 talks and the 11th coming soon....Look out for it.
- The attendees on Meetup range from 39 to 99.
- Speakers from Google, MS Research, Allen AI, Carneige Mellon university, UMass, and others.
- Some of our talk are available online.



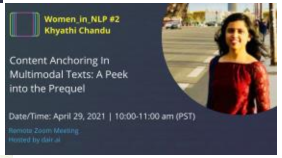


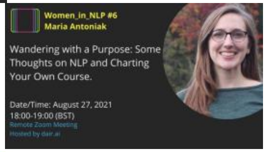


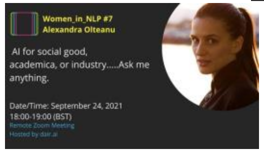


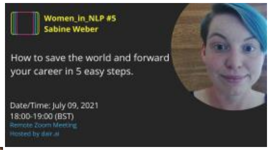
#Women_in_NLP

- The speakers share their latest research in NLP
  - To give the audience an idea of research directions in NLP.
- They also share their personal experience in NLP
  - Lesson learned.
  - Struggles.
  - Give advice internships, supervision, difference between research in academia and industry.



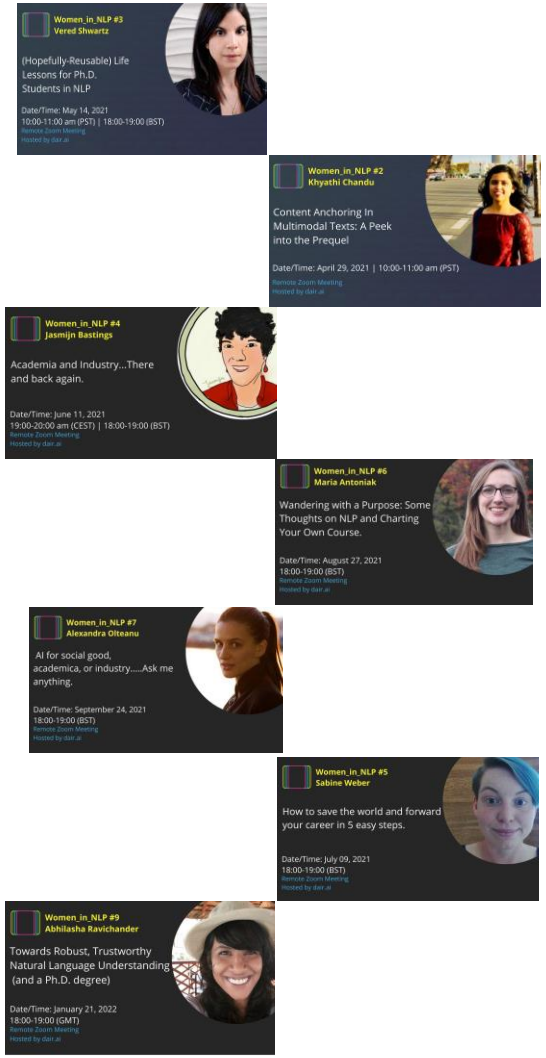






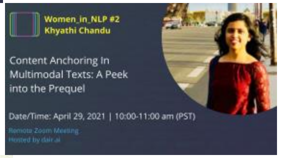
#Women_in_NLP

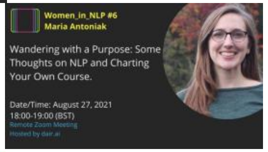
- Challenges:
    - Finding speakers.
    - Finding the right time.
    - Some events don’t continue.
    - Turn out is small.
    - Time and energy.

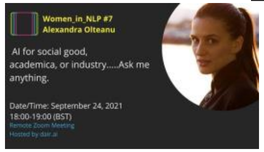


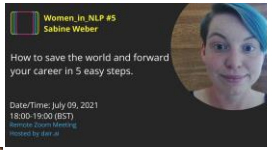
#Women_in_NLP

- Looking for co-organizers:
  - Contact me on e.fatma.e@gmail.com
  - Twitter @FatmaElsafoury
- To know about the latest talk:
  - Follow me on Twitter
  - Follow Meetup group https://www.meetup.com/dair-ai/









Thanks!

Fatma Elsafoury