Comparative Study on Word Embeddings and Social NLP Tasks

Fatma Elsafoury

Social media and cyberbullying

Grey social media platforms

[1] Emo, Love, and God: Making Sense of Urban Dictionary, a Crowd-Sourced Online Dictionary.
[2] Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board

Word embeddings

Social-Media-based

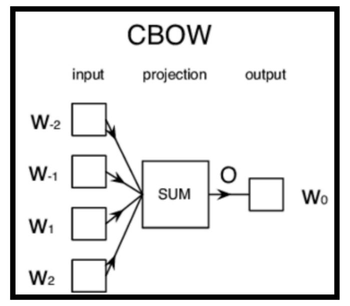
- Word embedding that are pre-trained on data collected from social media platforms.

200M tokens

fastText

UD

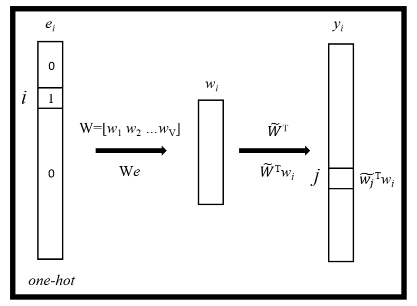
30M tokens



Chan

27B tokens

Glove model



Glove-Twitter

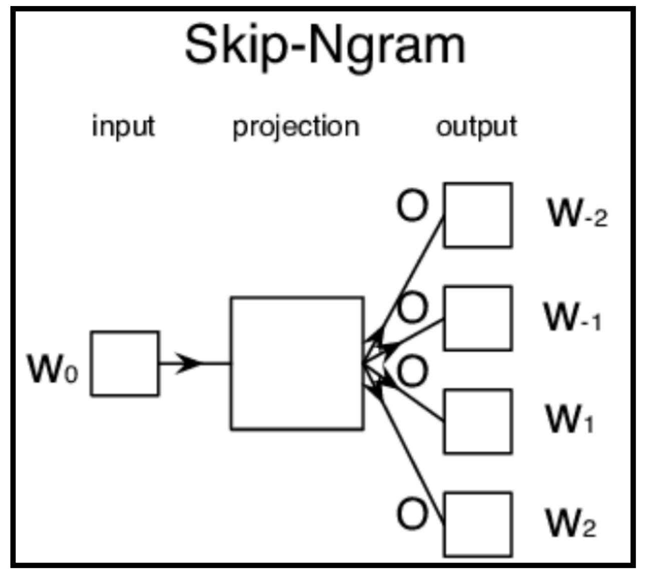
Word embeddings

Informational-based

- Word embeddings pre-trained on data collected from informational platforms like Google News or Wikipedia.

Google News

100B tokens

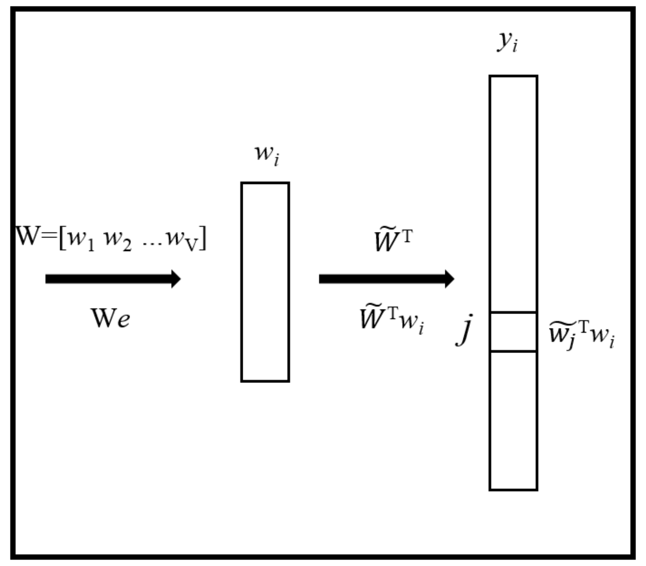


Word2vec

Wikipédia
L'encyclopédie libre

6B tokens

Glove model



Glove-WK

Social NLP tasks

Social-media-based vs. Informational-based

1. Cyberbullying detection:
   - Categorizing offenses.
   - Detecting cyberbullying in social media.

Word Embeddings	Similar words to "queer"
Word2vec	genderqueer, LGBTQ, gay, LGBT, lesbian
Glove-WK	transgender, lesbian, lgbt, lgbtq, bisexual
Glove-Twitter	fag, faggot, feminist, gay, cunt
Urban Dictionary	fag, homo, homosexual, bumblaster, buttyman
Chan	faggot, metrosexual, fag, transvestite, homo

Table1: The most similar 5 words to the word "queer"

Cyberbullying detection
Categorizing offenses

- Hurtlex lexicon:
  - 5963 offensive expression categorized in 11 groups

Category	Description
PS	ethnic slurs
IS	words related to social and economic disadvantage
QAS	descriptive words with potential negative connotations
CDS	derogatory words
RE	felonies and words related to crime and immoral behavior
PR	words related to prostitution
OM	words related to homosexuality
ASF	female genitalia
ASM	male genitalia
DDP	cognitive disabilities
DDF	physical disabilities

Table2: Hurtlext 11 offenses categories

Cyberbullying detection

Categorizing offenses

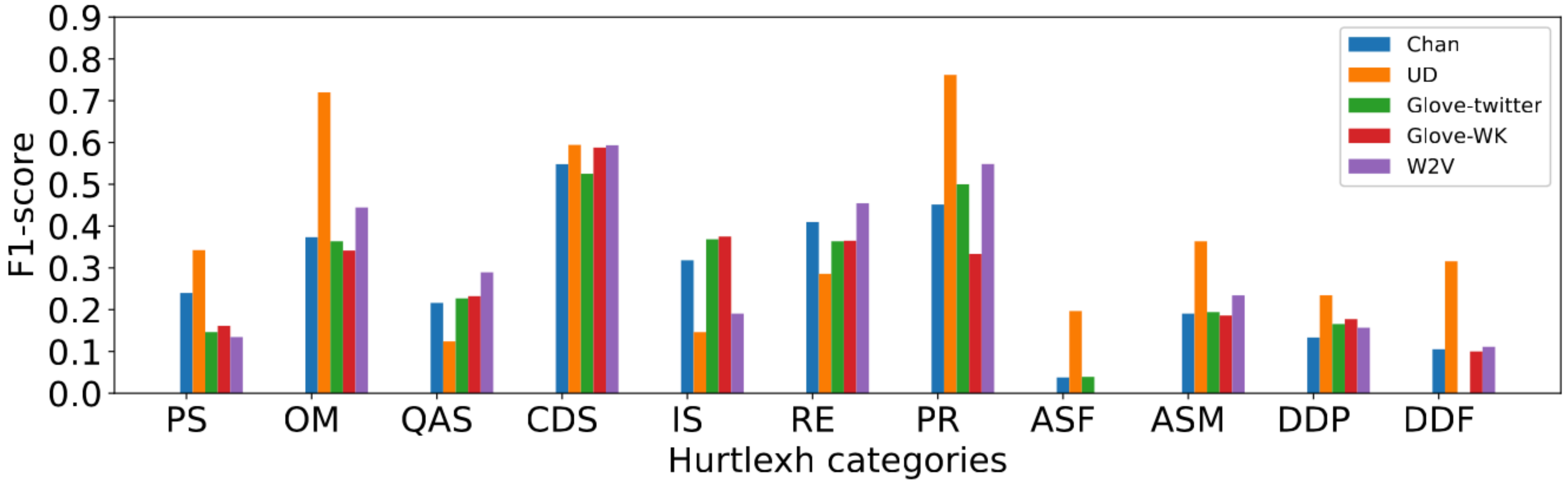


Figure 2: F1 scores of the KNN model with the different word embeddings on Hurtlext test set.

Cyberbullying detection

Categorizing offenses

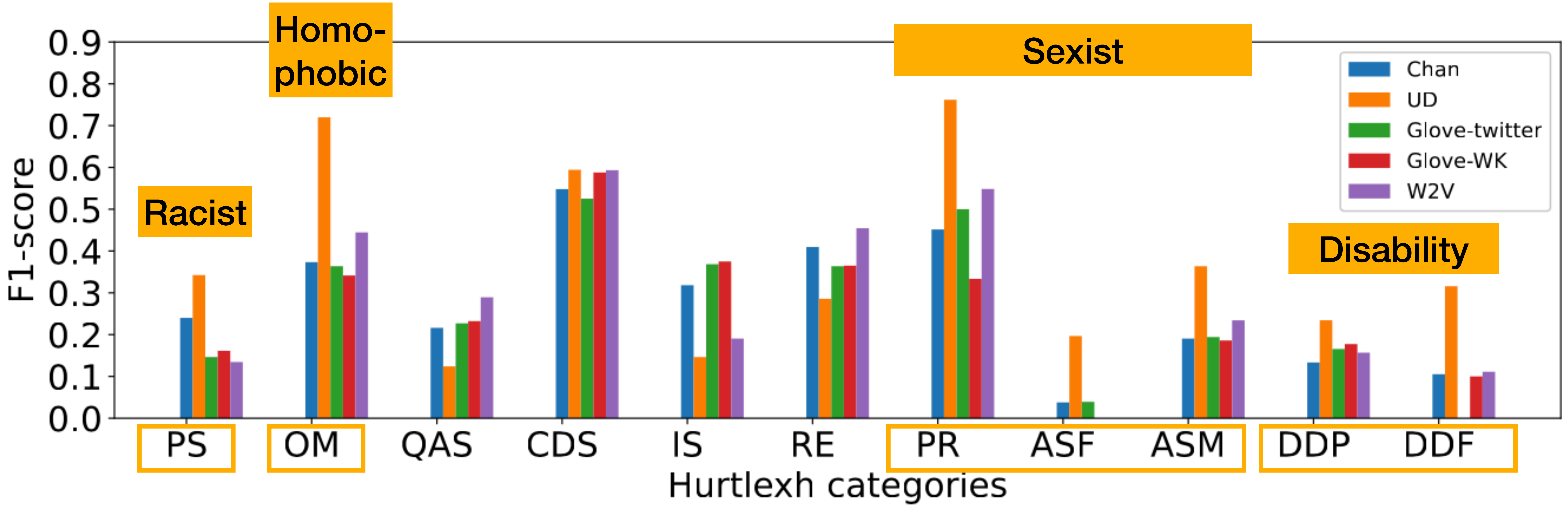


Figure 2: F1 scores of the KNN model with the different word embeddings on Hurtlext test set.

Cyberbullying detection

Categorizing offenses

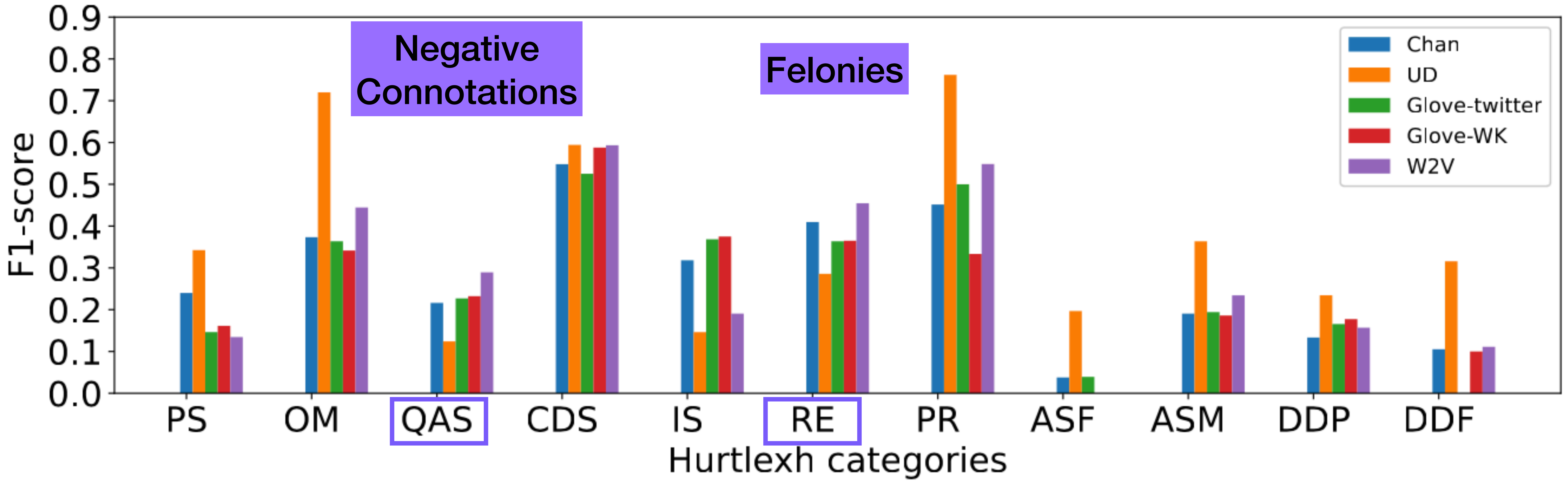


Figure 2: F1 scores of the KNN model with the different word embeddings on Hurtlext test set.

Cyberbullying detection

Categorizing offenses

- These results inspire two hypothesis:

(1) Cyberbullying-related Datasets

Social-media-based word embeddings UD, Chan, Glove-Twitter ↑ F1

Informational-based word embeddings: Word2vec, Glove-WK ↓ F1

(2) Cyberbullying-related Datasets

PS, PR, ASF, ASM,OM,SSP, DDF — UD ✓ F1

QAS, RE — W2v ✓ F1

Cyberbullying detection

Detecting cyberbullying in social media

- BiLSMT + Frozen embedding layer.

Dataset	Size	Pos.	Avg.	Max.
HateEval	12722	42%	21.75	93
Kaggle	7425	65%	25.28	1419
Twitter-sex	14742	23%	15.04	41
Twitter-rac	13349	15%	15.05	41
Jigsaw-tox	99738	6%	54	2321

Table 3: Cyberbullying-related datasets

Cyberbullying detection

Detecting cyberbullying in social media

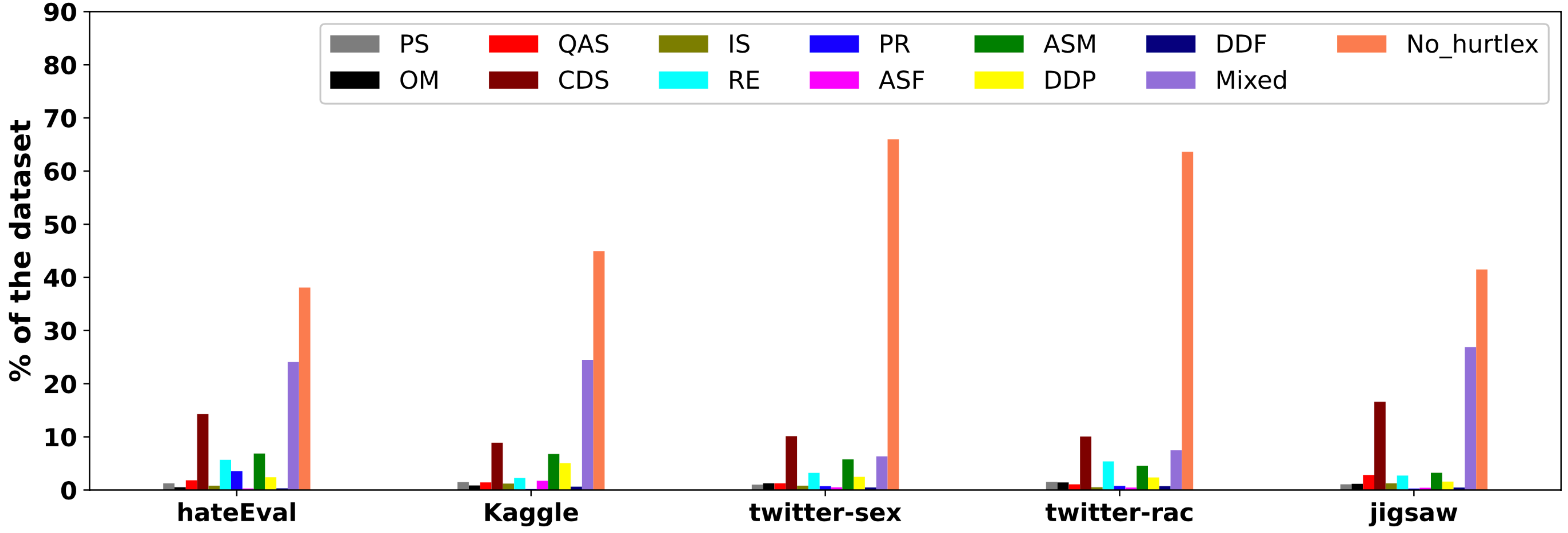


Figure 3: Percentage of each dataset that belong to the different Hurtlex categories

Cyberbullying detection

Findings

Social-media-based-word embeddings outperform Informational word embeddings

Certain word embeddings are better at detecting certain types of cyberbullying within our cyberbullying datasets

Table 4: The performance (F1 scores) of the BiLSTM model with each word embeddings
On the different Hurtlex category within our cyberbullying datasets

	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
HateEval														
Chan	0.615	0.444	0.615	**0.666**	0.555	**0.647**	0.658	0.421	0.555	**0.857**	0.5	0.570	0.730	0.602
UD	**0.7**	0.444	0.571	0.603	0.533	0.562	0.678	0.4	0.603	0.571	0.375	0.508	0.734	0.560
Glove-Twitter	0.695	**0.5**	**0.736**	0.663	0.631	0.619	**0.711**	0.620	0.690	0.571	0.285	**0.605**	**0.738**	**0.620**
Glove-WK	0.583	0.222	0.571	0.616	**0.666**	0.515	0.614	**0.72**	0.691	**0.857**	0.333	0.535	0.699	0.586
W2V	0.315	**0.5**	0.666	0.648	0.631	0.514	0.614	0.714	**0.72**	0.571	**0.666**	0.593	0.705	0.604
Kaggle														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.380	**0.777**	**1**	**0.760**	0.571	**0.545**	0.571	**1**	**0.666**	0.916	**0.909**	0.571	0.783	**0.727**
UD	**0.72**	0.761	**1**	0.703	**0.75**	0.461	0.75	0.666	0.507	0.888	0.8	**0.611**	**0.813**	0.725
Glove-Twitter	0.454	0.727	0.444	0.627	0.727	0.285	**0.823**	0	0.520	**0.923**	0.8	0.513	0.790	0.587
Glove-WK	0.5	0.625	**1**	0.588	0.666	0.5	0.666	0.666	0.507	0.869	0.666	0.525	0.8	0.660
W2V	0.352	0.375	**1**	0.602	0.25	0.4	0.714	**1**	0.526	0.818	0.666	0.479	0.797	0.614
Twitter-sexism														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.666	0.829	0.421	0.523	0.695	0.4	0.45	0.6	0.510	0.666	0.56	0.561	0.586	0.574
UD	0.666	0.8	0.521	0.656	0.75	0.510	0.608	**0.923**	0.622	**0.75**	**0.687**	0.629	0.695	0.678
Glove-Twitter	0.666	**0.863**	0.380	0.640	**0.8**	0.5	0.693	**0.923**	**0.653**	0.571	0.645	0.631	0.702	0.667
Glove-WK	0.666	0.818	**0.608**	**0.686**	0.740	**0.655**	0.734	0.727	0.636	**0.75**	0.685	**0.675**	0.708	**0.699**
W2V	**0.727**	0.772	0.571	0.598	0.695	0.56	**0.769**	0.833	0.623	**0.75**	0.666	0.650	**0.730**	0.688
Twitter-racism														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	**0.76**	0.736	0.8	0.732	0.5	**0.809**	**0.4**	0	0.428	0.588	**1**	0.671	0.784	0.631
UD	0.754	**0.956**	**0.909**	0.762	**0.6**	0.8	0.333	0	0.571	0.583	0.909	0.658	0.783	**0.663**
Glove-Twitter	0.72	0.8	**0.909**	0.734	0.5	0.790	**0.4**	0	**0.666**	0.636	0.909	**0.694**	**0.813**	0.659
Glove-WK	0.703	0.8	0.833	**0.784**	0.5	0.793	0.333	0	0.615	**0.761**	0.769	0.688	0.800	0.644
W2V	0.680	0.588	0.75	0.622	0.571	0.767	0.333	0	0.545	0.631	0.8	0.654	0.748	0.591
Jigsaw-Toxicity														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.15	0.45	**0.461**	0.427	**0.5**	0.310	0.285	0.75	0.652	0.553	0.482	0.484	0.658	0.474
UD	0.303	**0.615**	0.387	**0.441**	0.333	0.274	0.285	0.666	0.653	0.461	0.538	0.449	0.666	0.467
Glove-Twitter	0.285	0.578	0.322	0.433	0.444	0.360	0.444	**0.888**	**0.693**	0.553	**0.571**	**0.493**	**0.687**	**0.519**
Glove-WK	0.166	0.514	0.428	0.362	0.428	**0.407**	0.25	0.75	0.615	0.558	0.363	0.454	0.661	0.458
W2V	**0.333**	0.437	0.230	0.421	0.333	0.350	**0.545**	0.571	0.543	**0.588**	0.518	0.448	0.678	0.461

Social bias Analysis

Measuring bias

- Bias metrics: WEAT, RNSB, RND, ECT.
- Bias types: Gender and Racial bias.
- Hypothesis:

Social-media-based word embeddings:
UD, Chan, Glove-Twitter

↑ **Bias**

Informational-based word embeddings:
Word2vec, Glove-WK

↓ **Bias**

Measuring social bias

Results

	Gender Bias				**Racial Bias**			
Word embeddings	WEAT	RNSB	RND	ECT	WEAT	RNSB	RND	ECT
Word2vec	4 (0.778)	2 (0.033)	2 (0.087)	4 (0.752)	2 (0.179)	1 (0.095)	1 (0.151)	4 (0.786)
Glove-WK	**5 (0.893)**	4 (0.052)	4 (0.204)	2 (0.829)	**5 (0.439)**	2 (0.118)	4 (0.253)	1 (0.903)
Glove-Twitter	2 (0.407)	3 (0.041)	3 (0.127)	1 (0.935)	4 (0.275)	3 (0.122)	2 (0.179)	2 (0.898)
UD	1 (0.346)	1 (0.031)	1 (0.051)	**5 (0.652)**	1 (0.093)	4 (0.132)	3 (0.196)	**5 (0.726)**
Chan	3 (0.699)	**5 (0.059)**	**5 (1.666)**	3 (0.783)	3 (0.271)	**5 (0.299)**	**5 (2.572)**	3 (0.835)

Table 5: The Bias scores using the different metrics of the different word embeddings.

Other types of bias

- Most of the research focuses on gender and racial biases.
- Using slurs and third person profanity aims at stressing on the inferiority of the identity of the target of the attack [1].
- Since the internet and social media is rife with racial slurs and profanity, it is important to study how ML models encode this offensive stereotyping.

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

SOS: Systematic Offensive stereotype Bias

Definition

- A systematic association in the word embeddings between profanity and marginalized groups of people.
- NOI words.
- 15 word embeddings.

Group	Word
LGBTQ*	lesbian, gay, queer, homosexual, lgbt, lqbtq, bisexual, transgender, tran, non-binary
Women*	woman, female, girl, wife, sister, mother, daughter
Non-white ethnicities*	african, african american, black, asian, hispanic, latin, mexican, indian, arab, middle eastern
Straight	heterosexual, 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

SOS: Systematic Offensive stereotype Bias

NCSP: Normalized cosine similarity to profanity

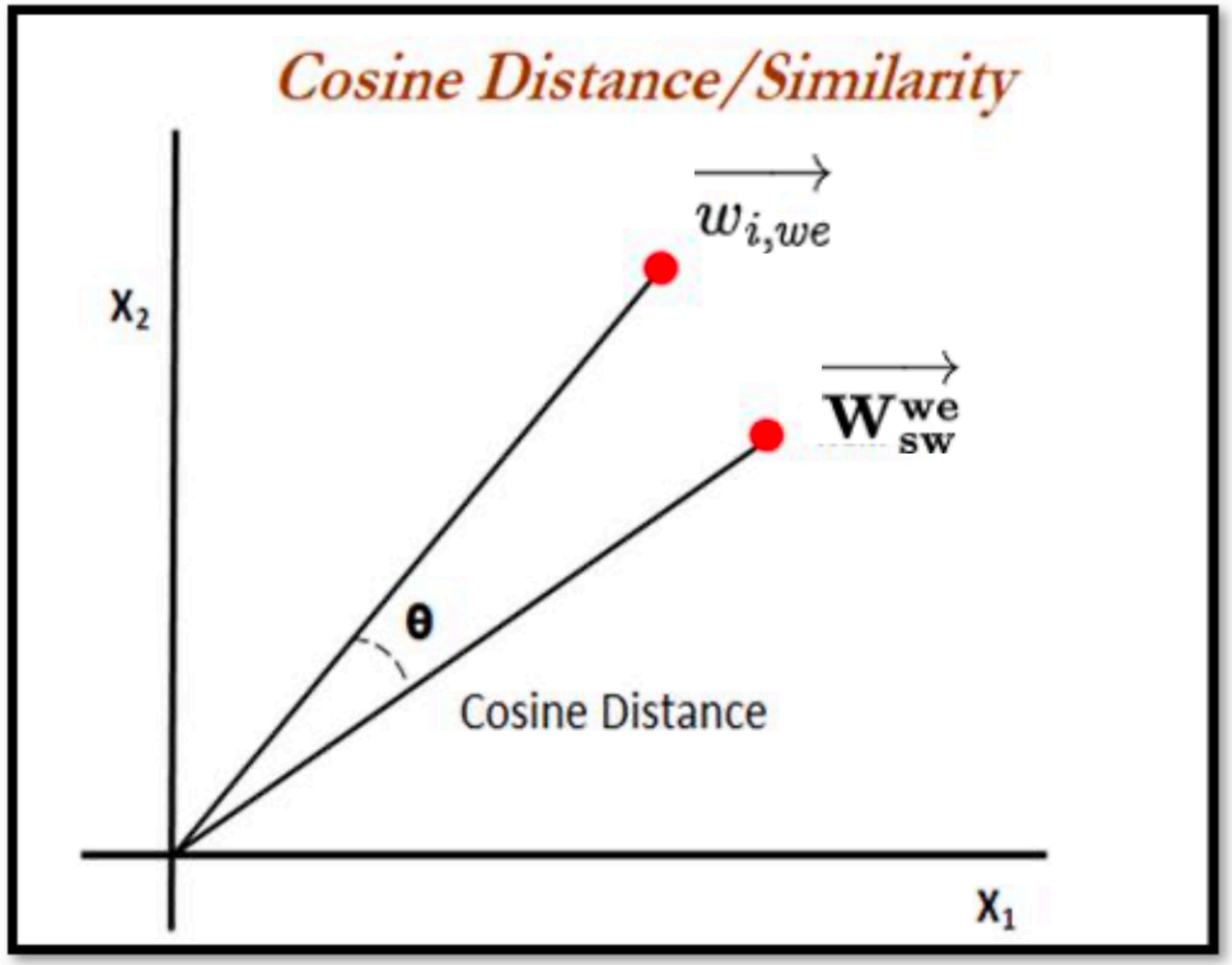
- Measure SOS Bias:

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

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

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



SOS: Systematic Offensive stereotype Bias
Results

Word embeddings	Mean SOS							
	Gender		Sexual orientation		Ethnicity		Marginalised vs. Non-marginalised	
	Women	Men	LGBTQ	Straight	Non-white	White	Marginalised	Non-marginalised
Word2Vec	**0.293**	0.209	0.475	**0.5**	**0.456**	0.390	**0.418**	0.340
Glove-WK	**0.435**	0.347	**0.669**	0.5	**0.234**	0.169	**0.464**	0.260
Glove-Twitter	**0.679**	0.447	**0.454**	0*	**0.464**	0.398	**0.520**	0.376
UD	**0.509**	0.436	**0.582**	0.361	**0.282**	0.244	**0.466**	0.319
Chan	**0.880**	0.699	**0.616**	0.414	**0.326**	0.176	**0.597**	0.373
Glove-CC	**0.567**	0.462	**0.480**	0.195	**0.446**	0.291	**0.493**	0.339
Glove-CC-large	**0.318**	0.192	**0.472**	0.302	**0.548**	0.278	**0.453**	0.252
FT-CC	**0.284**	0.215	0.503	**0.542**	**0.494**	0.311	**0.439**	0.301
FT-CC-sws	**0.473**	0.422	**0.445**	0.277	**0.531**	0.379	**0.480**	0.384
FT-Wiki	**0.528**	0.483	0.555	**0.762**	**0.393**	0.265	**0.496**	0.385
FT-Wiki-sws	0.684	0.684	0.656	**0.798**	0.555	**0.579**	0.632	**0.635**
SSWE	0.619	**0.651**	**0.438**	0*	**0.688**	0.560	**0.569**	0.537
Debias-W2v	**0.205**	0.204	0.446	**0.5**	**0.471**	0.420	**0.386**	0.356
P-DESIP	**0.266**	0.220	**0.615**	0.491	**0.354**	0.314	**0.434**	0.299
U-DESIP	**0.266**	0.220	**0.616**	0.492	**0.343**	0.299	**0.431**	0.283

*Glove-Twitter and SSWE did not include the NOI words that describe the "Straight" group.

Table 2: Mean SOS score of the different groups for all the word embeddings. Bold values represent the highest SOS score between the two different groups in each category (gender, sexual orientation, ethnicity, and marginalised vs. non marginalised).

SOS: Systematic Offensive stereotype Bias

Results

Most biased against LGBTQ

Most biased against Women

Most biased against Non-white ethnicity

Word embeddings	Mean SOS		
	Women	LGBTQ	Non-white
Word2vec	0.293	**0.475**	0.456
Glove-WK	0.435	**0.669**	0.234
glove-twitter	**0.679**	0.454	0.464
UD	0.509	**0.582**	0.282
Chan	**0.880**	0.616	0.326
Glove-CC	**0.567**	0.480	0.446
Glove-CC-large	0.318	0.472	**0.548**
FT-CC	0.284	**0.503**	0.494
FT-CC-sws	0.473	0.445	**0.531**
FT-WK	0.528	**0.555**	0.393
FT-WK-sws	**0.684**	0.656	0.555
SSWE	0.619	0.438	**0.688**
Debias-W2v	0.205	0.446	**0.471**
P-DESIP	0.266	**0.615**	0.354
U-DESIP	0.266	**0.616**	0.343

Table 3: The mean SOS bias score of each word embeddings towards each marginalised group. Bold scores reflect the group that the word embeddings is most biased against.

SOS: Systematic Offensive stereotype Bias

SOS vs social bias

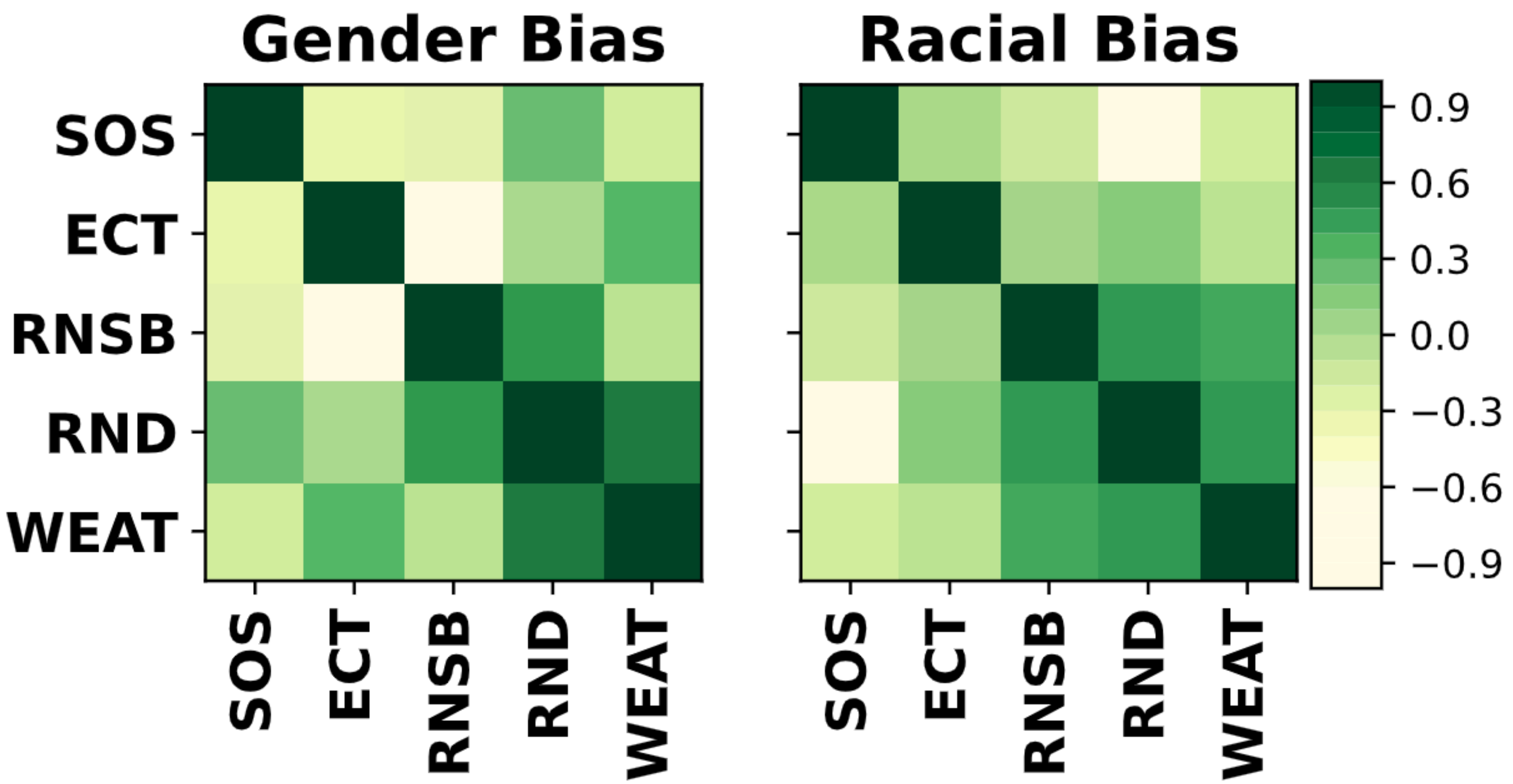


Figure 1: Spearman's correlation between the different bias metrics (SOS and social bias) for all the examined word embeddings. For gender bias, SOS refers to SOS_{women}, and for racial bias to $SOS_{non\text{-}white}$.

SOS: Systematic Offensive stereotype Bias

SOS Validation

- SOS vs. Online stats on online Hate (OEOH) in Germany, Finland, US, and UK. Most hateful content is targeted at, in order, LGBTQ, Non-white-ethnicities, and Women.

- NCSP vs. WEAT, RNSB, RND, and ECT to measure the SOS bias.

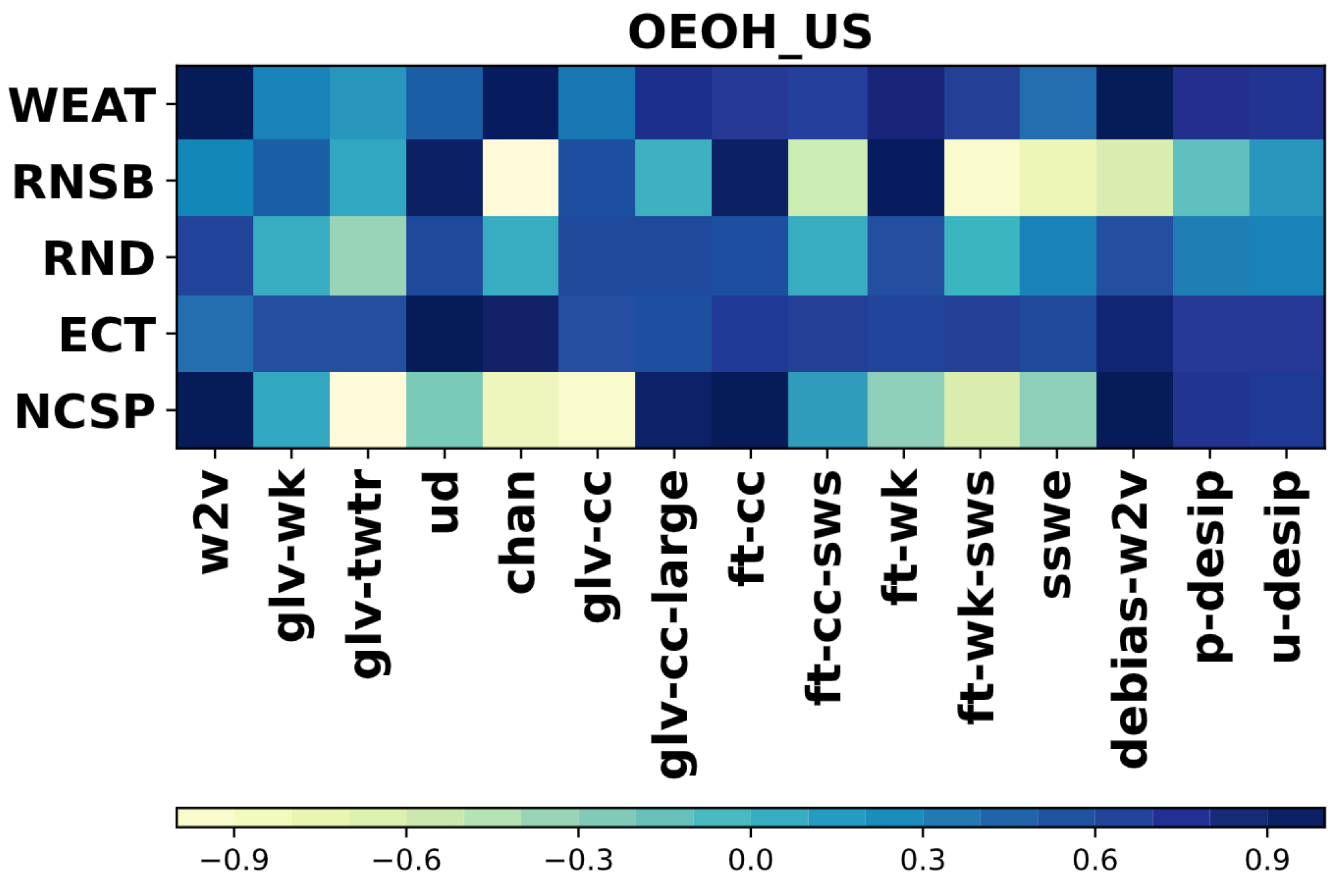


Figure 2: Pearson's correlation between the different SOS metrics and the percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online for the OEOH-US survey for the word embeddings.

The SOS bias influence on hate speech detection

Performance

Word embeddings	HateEval		Twitter-Hate		Twitter-racism		Twitter-sexism	
	MLP	BiLSTM	MLP	BiLSTM	MLP	BiLSTM	MLP	BiLSTM
Word2vec	0.593	0.663	0.681	0.772	0.683	0.717	0.587	0.628
Glove-WK	0.583	0.651	0.713	0.821	0.681	0.727	0.587	0.641
Glove-Twitter	0.623	0.671	0.775	0.851	0.680	0.699	0.589	**0.668**
UD	0.597	0.652	0.780	0.837	0.679	0.698	0.578	0.632
Chan	**0.627**	0.661	0.692	0.840	0.650	0.712	0.563	0.647
Glove-CC	0.625	**0.675**	0.778	0.839	0.695	0.740	0.577	0.648
Glove-CC-large	0.626	0.674	0.775	**0.860**	**0.709**	0.724	0.593	**0.668**
FT-CC	**0.627**	**0.675**	**0.792**	0.843	0.701	0.741	**0.607**	0.654
FT-CC-sws	0.605	0.660	0.746	0.830	0.701	**0.746**	0.588	0.657
FT-WK	0.606	0.650	0.784	0.827	0.699	0.706	0.601	0.653
FT-WK-sws	0.606	0.650	0.723	0.820	0.689	0.736	0.561	0.633
SSWE	0.558	0.628	0.502	0.715	0.324	0.666	0.171	0.548
Debiased-w2v	0.626	0.652	0.678	0.741	0.674	0.715	0.564	0.638
P-DESIP	0.575	0.657	0.697	0.817	0.673	0.731	0.538	0.650
U-DESIP	0.598	0.649	0.702	0.815	0.673	0.726	0.548	0.638

Table 5: F1 scores for the used models for hate speech detection using the examined word embeddings on the examined datasets. Bold values are the highest scores among the different word embeddings per model and dataset.

The bias influence on hate speech detection

Performance

Gender Bias

Dataset	Model	Spearman's correlation				
		WEAT	RNSB	RND	ECT	SOS_wmn
HateEval	MLP	**0.385**	-0.039	0.317	0	0.224
	BiLSTM	**0.303**	0.346	0.282	-0.214	0.064
Twitter-sexism	MLP	**0.732***	0.067	0.357	0.464	-0.214
	BiLSTM	0.042	-0.15	-0.117	-0.160	**0.107**
Twitter-hate	MLP	0.207	-0.157	0.042	0.15	**0.246**
	BiLSTM	**0.492**	0.117	0.421	0.028	0.446

Table 7: Spearman's rank correlation coefficient of the gender bias scores of the different word embeddings and the F1 scores of the used models for each bias metric and dataset. * describe the significant correlation with $p - value < 0.005$.

Racial Bias

Dataset	Model	Spearman's correlation				
		WEAT	RNSB	RND	ECT	SOS_eth
HateEval	MLP	-0.332	0.010	-0.228	-0.467	**0.285**
	BiLSTM	0.125	0.049	**0.228**	-0.110	0.096
Twitter-racism	MLP	-0.532*	-0.189	-0.142	-0.017	**0.132**
	BiLSTM	0.217	-0.057	**0.292**	-0.175	-0.392
Twitter-hate	MLP	-0.353	**-0.049**	-0.092	-0.278	-0.064
	BiLSTM	-0.175	0.060	0.028	-0.489	**0.185**

Table 8: Spearman 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.

The bias influence on hate speech detection

Fairness in downstream tasks (Extrinsic bias)

- Unfairness in ML in the case of Hate speech detection.
- g is marginalized groups.
- $\hat{g}$ is the non-marginalized groups.

$$Unfairness_{g,y} = FPR_g - FPR_{\hat{g}}$$

The bias influence on hate speech detection

Unfairness

Gender Bias

Dataset	Model	FPR gap				
		WEAT	RNSB	RND	ECT	SOS_wmn
HateEval	MLP	**0.196**	0.103	0.189	-0.16	-0.085
	BiLSTM	0.257	**0.382**	0.267	-0.178	-0.114
Twitter-sexism	MLP	**0.478**	0.271	0.278	0.075	0.053
	BiLSTM	-0.092	-0.282	-0.203	-0.167	-0.15
Twitter-hate	MLP	0.0285	**0.486**	0.439	0.067	0.287
	BiLSTM	-0.084	**0.384**	0.091	-0.016	-0.334

Table 9: Spearman correlation coefficient of the gender bias scores of the different word embeddings and the FPR gender gaps of the used models for each bias metric and dataset.

Racial Bias

Dataset	Model	FPR gap				
		WEAT	RNSB	RND	ECT	SOS _eth
HateEval	MLP	-0.092	-0.4	-0.273	-0.073	**0.507**
	BiLSTM	0.007	0.003	0.317	**0.535***	-0.199
Twitter-racism	MLP	-0.448	-0.204	-0.156	-0.037	-0.36
	BiLSTM	-0.242	0.174	-0.242	0.089	**0.247**
Twitter-hate	MLP	0.032	-0.107	-0.025	0.078	-0.143
	BiLSTM	-0.479	-0.458	-0.119	**0.21**	-0.028

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

Conclusion
Learned lessons

- Social-media-based word embeddings are better than informational based word embeddings on the task of offenses categorization and cyberbullying detection.
- Social-media based word embeddings are not significantly more socially biased than information word embeddings.
- All examined word embeddings contain SOS bias and most of them contain SOS bias against marginalized groups.
- There is no evidence that the bias (sos, gender, or racial) in the word embeddings has influence on the models's performance or fairness on the downstream tasks.

Conclusion

What is next?

- Understand how the bias influence downstream tasks in LLM.
- Study the influence of intrinsic and extrinsic debiasing methods on the downstream tasks in LLM.
- Learn where to focus our efforts to make LLM fairer: Upstream or downstream.