Systematic Offensive Stereotyping (SOS) Bias in Language Models

Fatma Elsafoury
Fraunhofer Research Institute
Weizenbaum Research Institute
fatma.elsafoury@fokus.fraunhofer.de

Abstract

In this paper, we propose a new metric to measure the SOS bias in language models (LMs). Then, we validate the SOS bias and investigate the effectiveness of removing it. Finally, we investigate the impact of the SOS bias in LMs on their performance and fairness on hate speech detection. Our results suggest that all the inspected LMs are SOS biased. And that the SOS bias is reflective of the online hate experienced by marginalized identities. The results indicate that using debias methods from the literature worsens the SOS bias in LMs for some sensitive attributes and improves it for others. Finally, Our results suggest that the SOS bias in the inspected LMs has an impact on their fairness of hate speech detection. However, there is no strong evidence that the SOS bias has an impact on the performance of hate speech detection.

1 Introduction

Language models (LMs) are the new state-of-the-art models. They are being implemented in tools like search engines (Zhu et al., 2023) and content moderation (Elsafoury et al., 2021). Research has shown that LMs, are socially biased (Nangia et al., 2020; Nadeem et al., 2021). However, the offensive stereotyping bias and toxicity in LMs are still understudied. Nozza et al. (2021) and Nozza et al. (2022) demonstrate that LMs tend to generate hurtful content. Ousidhoum et al. (2021) demonstrate that when probed by words that describe different identity groups, English LMs generate words that are insulting 24% of the time, in comparison to, stereotypical (13%), confusing (25%) and normal (38%). These results suggest that LMs are toxic. However, they do not investigate whether LMs systematically give higher probability to profane content over non-profane one, when probed by different identity groups.

On the other hand, Elsafoury et al. (2022), introduce systematic offensive stereotyping (SOS) bias and propose a method to measure it in 15 different static word embeddings. However, the SOS bias has not been measured in LMs. Moreover, Elsafoury et al. (2022) measure and validate the SOS bias, but their investigation does not include removing the SOS bias or how effective the state-of-the-art debias methods are on removing the SOS bias. Additionally, Elsafoury et al. (2022) investigate the impact of the SOS bias in static word embeddings on their performance on the task of hate speech detection, excluding the impact of the SOS bias on another critical aspect, which is the *fairness* of hate speech detection.

In this paper, we fill these research gaps by proposing a metric to measure the SOS bias in LMs (§3). Then, we validate the proposed SOS bias metric by comparing it to social bias metrics (§4.1). Additionally, we investigate how reflective the SOS bias is of the online hate experienced by marginalized groups (§4.2). Thereafter, we investigate the effectiveness of removing the SOS bias using one of the state-of-the-art debias methods (§5). Finally, we investigate the impact of the SOS bias in the inspected LMs on their performance (§6) and their fairness (§7) of the downstream task of hate speech detection. **The main Contributions** of this work can be summarized as following: **(1)** We provide a comprehensive investigation of the systematic offensive stereotyping (SOS) bias in LMs. **(2)** We create a new dataset to measure the SOS bias in LMs. **(3)** We make the newly created dataset and the code used in this work available online[1]

The **findings** of this work demonstrate that all the inspected LMs are SOS-biased. Our results suggest that for most of the examined sensitive attributes, the SOS bias scores are higher against marginalized identities. We demonstrate that the SOS bias in the inspected LMs is reflective of the hate and extremism that are experienced by

[1]This link will be available upon acceptance.

marginalized groups online. However, we found no strong evidence that our proposed SOS bias metric reveals different information from social bias metrics. Our results suggest that removing SOS bias from LMs, using one of the state-of-the-art debias methods, improved the SOS bias scores in the inspected LMs regarding some sensitive attributes and worsened it for others. On the other hand, our results suggest that, for some bias metrics, removing the SOS bias significantly improved the social bias scores. Our results suggest that the SOS bias in LMs has an impact on their fairness on hate speech detection. However, there is no strong evidence that the SOS bias has an impact on the performance of the task of hate speech detection, which is inline with previous findings (Goldfarb-Tarrant et al., 2021).

2 Background

There are a few studies that investigate the toxic response of LMs when probed by different identity groups. Nozza et al. (2021) investigate the hurtful stereotypical text generated by LMs when prompted by template sentences that contain gendered identity words. The results show that when LMs prompted by female gendered identity, 9% of the generated text referred to sexual promiscuity. Nozza et al. (2022) follow similar approach to measure the hurtful stereotyping in sentence completion against the LGBTQIA+ community and found that 13% of the time, LMs generated identity attacks. Ousidhoum et al. (2021) demonstrates that LMs are toxic against people from different communities, marginalized and non-marginalized. The authors use the masked language models (MLM) task to predict words corresponding to template sentences that contain words that describe different identity groups, and then the authors use a logistic regression model to label whether the predicted words are toxic or not. Finally, a human evaluation of a 100 of the predicted words was conducted, where the results indicate that only 24% of the predicted words were insulting regardless of the context in the English LMs, 11% in French LMs, and 12% in Arabic LMs.

Even though these studies show evidence that LMs are toxic, especially towards marginalized groups. They have some limitations. For example, they do not systematically measure the toxicity or offensive stereotyping in LMs. As they all rely on open text generation by the LMs, which could be normal, confusing, or hurtful (Ousidhoum et al., 2021). We speculate that this is the reason behind the low percentages of hurtful content that are being exposed by these studies. Moreover, Ousidhoum et al. (2021) use a logistic regression model to predict whether the generated text is toxic or not and then uses human annotators to verify the label. This method of measuring the toxicity in LMs is not sustainable, as human annotators could be biased (Shah et al., 2020) and not always accessible. Elsafoury et al. (2022) introduce systemic offensive stereotyping (SOS) bias and propose a method to measure in static word embeddings. However, the SOS bias has not been yet investigated in LMs.

On the other hand, there are various metrics in the literature to systematically measure social bias in LMs like *SEAT* (May et al., 2019), *CrowS-Pairs* (Nangia et al., 2020), and *StereoSet* (Nadeem et al., 2021). In SEAT, the authors, inspired by the WEAT metric (Caliskan et al., 2017) to measure bias in static word embeddings, propose a method to measure social bias in LMs. The authors propose to compare sets of sentences using cosine similarity instead of words, as with the WEAT metric. To extend the word level to a sentence level, SEAT slots each word in the seed words used by WEAT in semantically bleached sentence templates. Similarly, *CrowS-Pairs* and *StereoSet* metrics are used to measure social bias in LMs. But instead of sentence templates, the authors use crowdsourced sentences and the MLM task to measure the social bias. The Crows-Pairs dataset contains 1,508 sentence pairs (stereotypical and non-stereotypical) and measures nine types of social bias. The StereoSet dataset contains 8,498 sentence pairs to measure four types of social bias.

To measure the systematic offensive stereotyping (SOS) bias in LMs, these metrics will fall short since the crowdsourced sentences contain socially stereotypical versus non-stereotypical sentences. In this paper, we mitigate the limitations of the current literature by proposing a method to measure SOS bias in LMs. We build on existing social bias metrics but instead of using stereotypical and non-stereotypical sentence-pairs, we create a new dataset of profane and non-profane sentence-pairs to measure SOS bias.

3 Measure SOS bias in LMs

The SOS bias, as defined by Elsafoury (2022) is, "*A systematic association in the word embeddings*

between profanity and marginalized groups of people". We use that definition to measure the SOS bias in three LMs, BERT-base-uncased (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), and ALBERT-base (Lan et al., 2020). To measure the SOS bias in LMs, we draw inspiration from the CrowS-Pairs and metric (Nangia et al., 2020) that uses the MLM task to measure social bias in LMs. We use the MLM task to measure how many times a LM associates a profane sentence versus a non-profane sentence with a certain identity groups.

3.1 SOS_{LM} bias dataset

To measure the SOS bias in LMs using the MLM task, we first need a dataset of sentence-pairs similar to the datasets used in the CrowS-Pairs and StereoSet metrics. But that dataset should have two key characteristics: 1) It should contain profane and non-profane sentence-pairs, instead of social stereotypical and non-stereotypical sentence-pairs. 2) It should also contain fine-grained information on the identity group present in the sentences, so we can measure the SOS bias against marginalized and non-marginalized identity groups.

We create a dataset that contains these two characters using sentence templates, similar to work done in the literature (May et al., 2019; Ousidhoum et al., 2021; Nozza et al., 2022). We use non-offensive identity (NOI) words to describe different identity groups Elsafoury et al. (2022). Because, as argued by Antoniak and Mimno (2021), NOI words are more coherent in describing different identity groups. Furthermore, we compiled a list of 78 NOI words to describe different marginalized and non-marginalized identities for six sensitive attributes: gender, race, sexual orientation, religion, disability[1], and social class (Elsafoury et al., 2022; Dixon et al., 2018; Sweeney and Najafian, 2019; Borkan et al., 2019; Nangia et al., 2020; Nadeem et al., 2021). The NOI words are described in Appendix A (Table 11).

Then, to create a list of profane and non-profane sentence-pairs, we need a list of profane and non-profane words. We use GPT3 model, on the OpenAI platform[2], to automatically generate profane and non-profane words. We used the following prompts[3]: *1) Write a list of offensive words* and *2) Write a list of nice words*. GPT3 generated 21 profane and 21 non-profane words, which are then used with sentence templates to create the synthesized dataset. Using the NOI words and the profane and non-profane word lists, we create a synthesized dataset that meets our criteria. Examples of the template sentence-pairs are provided in Table 1. The final synthesized dataset contains 1638 sentence-pairs.

Profane sentence	Non-profane sentence	Identity
You are a dumb	You are a friendly	Asian
You are a moronic	You are a incredible	Asian
You are a fucked	You are a caring	Asian
You are a awful	You are a best	Asian
....	....	.....
You are a dumb	You are a friendly	Dutch
You are a moronic	You are a incredible	Dutch
You are a fucked	You are a caring	Dutch
You are a awful	You are a best	Dutch

Table 1: Examples of template profane/non-profane sentence-pairs. The identity column contains NOI words to describe both marginalized and non-marginalized identities to fill in the blanks in the profane and the non-profane sentences.

3.2 SOS_{LM} bias metric

To measure the SOS bias in LMs, we use the MLM task. For a profane sentence (S) where, $S = U \cup M$, U is a set of unmodified tokens for example, $U = \{you, are, a, arab\}$ with length $|C|$, and M is a set of modified tokens for example, $M = \{vile\}$. To estimate the probability of the unmodified token conditioned on the modified tokens $p(U|M,\theta)$, we use the *pseudo-log-likelihood* (Nangia et al., 2020). The profane sentence $score(S)$ is then measured as:

$$score(S) = \sum_{i=0}^{|C|} logP(u_i \in U|M,\theta) \quad (1)$$

The same score is also measured for the non-profane sentence (S') where $S' = U \cup M'$, U is a set of unmodified tokens for example, $U = \{you, are, a, arab\}$ with length $|C|$, and M' is a set of modified tokens for example, $M' = \{nice\}$.

$$score(S') = \sum_{i=0}^{|C|} logP(u_i \in U|M',\theta) \quad (2)$$

Then, the bias scores are measured as the percentage of examples where the model (θ) assigns a higher probability estimate to the profane

[1]We use only words to describe disability because the words that describe the able-bodied are not commonly used and are not shared in the reviewed literature that is used to compile that list of words.

[2]https://platform.openai.com/overview

[3]We generated these words back in summer 2022. We acknowledge that the same prompt might generate different words.

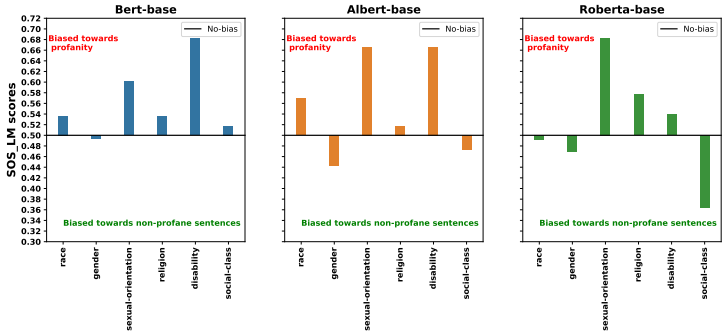

Figure 1: SOS_{LM} bias scores in the different LMs against all identity groups (marginalized and non-marginalized).

	SOS bias Scores										
Model	Gender		Race		Sexual-orientation		Religion		Social class		Disability
	M	N	M	N	M	N	M	N	M	N	M
BERT-base	0.476	**0.510**	**0.580**	0.501	0.576	**0.714**	0.523	**0.555**	**0.560**	0.480	**0.682**
AlBERT-base	**0.448**	0.435	0.542	**0.589**	**0.671**	0.642	0.495	**0.555**	**0.492**	0.457	**0.666**
RoBERTa-base	**0.517**	0.421	**0.519**	0.472	0.666	**0.761**	0.561	**0.603**	**0.391**	0.338	**0.539**

Table 2: SOS_{LM} scores of the different identity groups for all the language models. Bold values represent higher SOS bias scores between the marginalized (M) and the non-marginalized (N) groups in each sensitive attribute.

sentences (S) over the non-profane sentences (S') as in equation 3 where (N) is the number of sentence-pairs. If the percentage is over or below 0.5, then that means the model prefers profane or non-profane sentences, respectively, and is biased. On the other hand, if the percentage is 0.5, that means the model randomly assigns probability and hence is not biased. Since the focus of this paper is to measure the offensive stereotyping bias, we only consider a LM to be SOS-biased, if the $SOS_{LM} > 0.5$.

$$SOS_{LM} = \frac{Count(score(S) > Score(S'))}{N} \quad (3)$$

3.3 SOS biased LMs

We first measure the SOS bias scores against all the identity groups, marginalized and non-marginalized. The measured SOS bias scores in Figure 1 show that the majority of the inspected LMs are SOS biased, with ($SOS_{LM} > 0.5$), for the following sensitive attributes: race, sexual-orientation, religion, and disability. This indicates that the inspected LMs, in general, prefer profane sentences to non-profane ones. Then, we inspect the results closely to investigate whether the SOS bias scores in the inspected LMs are higher against the marginalized identity groups. We measure the SOS bias scores for the marginalized groups (M) and the non-marginalized groups (N) separately. Then, we compare the SOS bias scores between the marginalized and the non-marginalized identities. The results in Table 2 show that the majority of

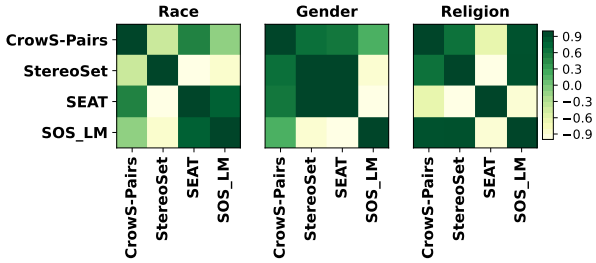

Figure 2: Heatmap of the Pearson's correlation (ρ) between the SOS_{LM} bias and social bias scores.

the models have higher bias scores against the marginalized identity groups for the following sensitive attributes: gender, race, social class, and disability (not statistical significant difference at $\alpha = 0.05$). We speculate that the higher SOS bias scores against marginalized groups could be a result of using biased pre-training datasets and an optimization method that might exacerbate that bias, as discussed by Shah et al. (2020).

On the other hand, the majority of the models have higher SOS bias scores against the non-marginalized groups for the sexual-orientation and religion sensitive attributes (no statistical significant difference at $\alpha = 0.05$). We speculate that this is the case because LMs might be SOS biased against the attribute itself. In other words, LMs consider these topics to be taboos and associate profanity with any mention of any sexual orientation or religion, marginalized or not.

CrowS-Pairs			
Bias	BERT	RoBERTa	AlBERT
Gender	0.580	**0.606**	0.541
Race	**0.581**	0.527	0.513
Religion	0.714	**0.771**	0.590
StereoSet			
Bias	BERT	RoBERTa	AlBERT
Gender	0.602	**0.663**	0.599
Race	0.570	**0.616**	0.575
Religion	0.597	**0.642**	0.603
SEAT			
Bias	BERT	RoBERTa	AlBERT
Gender	0.620	**0.939**	0.622
Race	**0.620**	0.307	0.551
Religion	**0.491**	0.126	0.430

Table 3: Social bias scores in LMs. **Bold** scores mean higher bias scores and more biased models.

4 SOS bias validation

We validate two aspects of the SOS_{LM} bias metric. The first aspect is how different it is from social bias metrics proposed in the literature. The second aspect is how reflective it is of the online hate experienced by marginalized identity groups.

4.1 SOS bias vs. social bias in LMs

We investigate the difference between the measured SOS bias scores and the social bias scores in the inspected LMs. We first measure the social bias scores in the LMs (Bert-base-uncased, AlBERT-base, and RoBERTa-base) using three bias metrics CrowS-Pairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), and SEAT (May et al., 2019). The social bias scores are reported in Table 3.

To investigate the difference between social bias and SOS bias scores, we measure the Pearson correlation coefficient (ρ) between the SOS bias scores measured using the proposed SOS_{LM} metric and the social bias scores measured using CrowS-Pairs, StereoSet, and SEAT metrics. The correlation is measured for three sensitive attributes: race, gender, and religion, as these attributes are common among all the used social bias metrics. Figure 2 shows that, there is a positive correlation between the measured SOS bias scores and social bias scores measured using different bias metrics. However, the positive correlation is not consistent across the different sensitive attributes. The most consistent positive correlation is found between the SOS bias scores and the Crows-Pairs scores. This could be because our SOS_{LM} metric uses a similar method to the CrowS-Pairs metric to measure SOS bias. These results suggest that, unlike the case with static word embeddings (Elsafoury et al., 2022), our proposed metric to measure the SOS bias in LMs does not reveal different information from that revealed by social bias metrics, especially when measured using the CrowS-Pairs metric.

Country	Sample size	Ethnicity	LGBTQ	Women
Finland	555	0.67	0.63	0.25
US	1033	0.6	0.61	0.44
Germany	978	0.48	0.5	0.2
UK	999	0.57	0.55	0.44

Table 4: The percentage of examined marginalized groups that experience online hate and extremism (Hawdon et al., 2015)

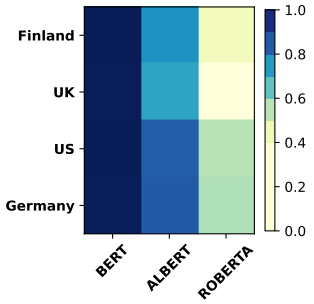

Figure 3: Heat-map of the Pearson's correlation (ρ) between the SOS bias scores measured using the SOS_{LM} metric and the percentages of marginalized identities who experience online hate in different countries.

4.2 SOS bias and online hate

We investigate how reflective the SOS bias is, in the inspected LMs against marginalized identity groups, of the online hate that the same marginalized groups experience. We use published statistics of the percentages of marginalized groups that experience online hate and extremism (Hawdon et al., 2015). Table 4 reports these statistics. We measure the Pearson correlation coefficients (ρ) between the online hate statistics and the SOS bias scores measured using our proposed SOS_{LM} metric against the marginalized groups (M) in Table 2 for the following sensitive attributes: race, gender, and sexual-orientation.

The results in Figure 3, show a strong positive correlation between the SOS bias measured in the inspected LMs using the proposed SOS_{LM} metric and the published percentages of marginalized people who experience online hate and extremism in Finland, Germany, the US, and the UK. This strong positive correlation exists for BERT, followed by AlBERT and then RoBERTa. These results suggest that the proposed metric of measuring SOS bias in LMs is reflective of the hate

Model	Crows-Pairs			StereoSet			SEAT			SOS_{LM}		
	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion
AlBERT-base	0.541	0.513	0.590	0.599	0.575	0.603	0.622	0.551	0.430	0.440	0.570	0.520
+ SentDebias-SOS	↓0.503	↑0.743	↑0.714	↓0.504	↓0.468	↓0.539	0.622	0.551	0.430	↑0.639	↓0.504	↓0.485
BERT-base-uncased	0.580	0.581	0.714	0.607	0.570	0.597	0.620	0.620	0.491	0.490	0.540	0.540
+ SentDebias-SOS	↓0.572	↓0.473	↓0.609	↓0.485	↓0.430	↓0.436	0.620	0.620	0.491	↑0.782	↑0.581	↓0.361
RoBERTa-base	0.606	0.527	0.771	0.663	0.616	0.642	0.939	0.307	0.126	0.470	0.490	0.580
+ SentDebias-SOS	↓0.494	↑0.567	↓0.361	↓0.517	↓0.463	↓0.457	0.939	0.307	0.126	↑0.734	↓0.285	↓0.438

Table 5: Social and SOS bias scores in the different models using different bias metrics before and after removing SOS bias using SentDebias. (↑) means that the bias score increased and the bias in the LMs worsened. (↓) means that the bias score decreased, and the bias in the LMs improved. The SOS bias scores reported here are against identity groups marginalized and non-marginalized.

that women, non-white ethnicities, and LGBTQ communities experience online. These results are inline with previous research on the SOS bias in static word embeddings (Elsafoury et al., 2022).

5 SOS bias removal

We investigate the effectiveness of one of the state-of-the-art social bias removal methods in the literature, on removing the SOS bias in LMs. We use SentDebias (Liang et al., 2020) to remove different types of bias from the LMs by projecting a sentence representation onto the estimated bias subspace and subtracting the resulting projection from the original sentence representation. Liang et al. (2020) compute the bias subspace by following these steps: 1) Define a list of identity words, e.g., "woman/man"; 2) Contextualize the identity words into sentences by finding sentences that contain those identity words in public datasets like SST[1] and WikiText-2[2]; 3) Obtain the representation of the contextualized sentence from the pre-trained LM; 4) Principal Component Analysis (PCA) (Abdi and Williams, 2010) then used to estimate principal directions of variations of the sentences' representations. The first K principal components are taken to define the bias subspace. For removing the SOS bias, we follow the same debias steps as (Liang et al., 2020). But instead of computing the social bias subspace, we compute the profanity subspace. We first contextualize the 21 profane and 21 non-profane words used in section 3.1 using the WikiText-2 dataset as explained earlier. Then we remove the profanity subspace from the inspected LMs. We build on the implementation shared by (Meade et al., 2022) to remove gender, racial, religion, and SOS bias.

The results, in Table 5, show that, in some cases, removing the SOS bias improved the social bias scores according to CrowS-Pairs and StereoSet. On the other hand, according SOS_{LM}, SentDebias improved the SOS bias scores for some sensitive attributes (race and religion) and worsened the bias in other sensitives attributes (gender). We found the same results for the SOS bias scores when measured against all identity groups, SOS bias scores measured against marginalized identities (M) and SOS bias scores measured against non-marginalized identities.The SEAT metric, did not show any difference in the bias scores for the debiased models, unlike the reported scores in (Meade et al., 2022).

Then, we calculate the T-test statistical significance test between the two independent samples of bias scores before and after applying the SentDebias algorithm to remove the SOS bias. We use the bias scores as measured by the CrowS-Pairs, StereoSet and SOS_{LM} metrics since these are the metrics that have different results after removing the SOS bias. The results show that according to StereoSet removing the SOS bias significantly improved the social bias scores for AlBERT ($pvalues = 0.01$), BERT ($p - value = 0.002$), and RoBERTa ($pvalue = 0.002$) at $\alpha = 0.05$.

So far, we introduced the SOS_{LM} metric to measure the SOS bias in LMs, validated it, and investigated the effectiveness of its removal. In the rest of this paper, we investigate the impact of the SOS bias in LMs on the performance and fairness of the downstream task of hate speech detection.

6 Impact of SOS bias on the performance of hate speech detection

To evaluate the performance of the inspected LMs on the task of hate speech detection, we first fine-tune the inspected LMs on the hate speech related datasets described in Table 6.

We follow the same pre-processing steps

[1] https://huggingface.co/datasets/sst

[2] https://huggingface.co/datasets/wikitext

Dataset	Samples	Positive samples	
Twitter-sexism	14742	23%	(Waseem and Hovy, 2016)
Twitter-racism	13349	15%	(Waseem and Hovy, 2016)
Civil-community	426707	0.08%	(Elsafoury et al., 2023)
Kaggle-insults	7425	35%	(Kaggle, 2012)
WTP-agg	114649	13%	(Wulczyn et al., 2017)
WTP-tox	157671	10%	(Wulczyn et al., 2017)

Note: Positive samples refer to offensive comments

Table 6: Statistics of hate speech datasets used with the inspected language models.

Dataset	BERT	AlBERT	ROBERTA
Kaggle	0.844	0.832	**0.847**
Twitter-sexism	0.871	**0.884**	0.880
Twitter-racism	**0.930**	0.924	0.929
WTP-agg	0.937	**0.939**	0.934
WTP-tox	0.960	0.961	**0.963**
Civil-community	0.582	0.558	**0.589**

Table 7: F1 scores of the inspected LMs on the different hate speech datasets. **Bold** values denote the best performance.

described in (Elsafoury et al., 2021), as the authors fine-tune BERT on the task of hate speech detection, which are: (1) remove URLs, user mentions, non-ASCII characters, and the retweet abbreviation "RT" (Twitter datasets). (2) All letters are lower cased. (3) Contractions are converted to their formal format. (4) A space is added between words and punctuation marks. We then split the datasets into 40% training set, 30% validation set and 30% test set. We train the models for 3 epochs, using a batch size of 32, a learning rate of $2e^{-5}$, and a maximum text length of 61 tokens. The performance results (F1-scores) are reported in Table 7. Then, to investigate the impact of the SOS bias in the LMs on their performance on hate speech detection, we use correlation. We compute the Pearson correlation coefficient (ρ) between the F1-scores of the LMs reported in Table 7 and the SOS bias scores against the marginalized identities (M) displayed in Table 2. The results, in Table 8, show a strong positive correlation with the SOS bias scores against marginalized identities in all the datasets: Twitter-racism (race, gender, and religion); WTP-agg (race, disability, and social class); and WTP-toxicity (gender, sexual-orientation, and religion); Kaggle (gender, religion); Civil-community (gender, and religion); and Twitter-sexism (sexual-orientation). However, these results are not consistent across all the sensitive attributes. We speculate that this is due to the different targets of the hate in the different hate speech datasets. For example, the SOS bias scores in the gender, race, sexual-orientation, and religion

	Sensitive attribute					
Dataset	Race	Gender	Sexuality	Religion	Disability	Social class
Kaggle	-0.049	0.903	-0.371	0.912	-0.574	-0.297
Twitter-sexism	-0.772	-0.195	0.966	-0.216	-0.315	-0.589
Twitter-racism	0.292	0.705	-0.664	0.719	-0.262	0.043
WTP-agg	0.477	-0.999	-0.068	-0.999	0.872	0.682
WTP-toxicity	-0.945	0.732	0.724	0.718	-0.973	-0.996
Civil-community	-0.075	0.915	-0.346	0.923	-0.595	-0.323

Table 8: Pearson Correlation Coefficient (ρ) between the SOS bias scores against the marginalized groups in the inspected LMs and the F1 scores of the different LMs on each dataset.

sensitive attributes, correlate positively with the performance of the LMs on the following hate speech dataset: Twitter-racism, WTP-aggression, and WTP-toxicity where the targets of the hate in the datasets are matching the marginalized groups in the gender, race, sexual-orientation, and religion sensitive attributes. Yet, these results and the impact of the SOS bias on the performance of hate speech detection remain inconclusive.

7 Impact of SOS bias on fairness of hate speech detection

To measure the impact of the SOS bias in the inspected LMs on their fairness of the task of hate speech detection, we first need to measure the fairness of the inspected LMs on hate speech detection. To this end, we use the fairness scores reported in (Elsafoury et al., 2023). Where the authors measure fairness as the absolute difference in the false positive rates (FPR), true positive rates (TPR), and the area under the curve (AUC) between the marginalized group (g) and non-marginalized group ($\hat{g}$), as shown in eq. (4), eq. (5), and eq. (6). These scores measure the unfairness of the model in how it treats different identity groups of people differently. The higher the score, the more unfair the model is and the lower the scores, the better the model in terms of fairness. Table 9 describes the inspected identity groups in three sensitive attributes: gender, race, and religion. Elsafoury et al. (2023) measure fairness of the same LMs that we use in this work, AlBERT-base, BERT-base, and RoBERTa-base, on the downstream task of hate speech detection using the Civil comments dataset described in Table 6.

$$FPR\_gap_{g,\hat{g}} = |FPR_g - FPR_{\hat{g}}| \quad (4)$$

$$TPR\_gap_{g,\hat{g}} = |TPR_g - TPR_{\hat{g}}| \quad (5)$$

$$AUC\_gap_{g,\hat{g}} = |AUC_g - AUC_{\hat{g}}| \quad (6)$$

The fairness scores reported in Table 10 show that

Sensitive attribute	Marginalized	Non-marginalized
Gender	Female	Male
Race	Black and Asian	White
Religion	Jewish and Muslim	Christian

Table 9: The inspected identity groups to measure fairness.

Attribute	Model	FPR_gap	TPR_gap	AUC_gap
Gender	AlBERT	0.006	0.038	0.003
	BERT	0.008	0.036	0.009
	RoBERTa	0.004	0.031	0.011
Race	AlBERT	0.008	0.001	0.018
	BERT	0.015	0.002	0.025
	RoBERTa	0.003	0.011	0.021
Religion	AlBERT	0.009	0.108	0.020
	BERT	0.008	0.062	0.012
	RoBERTa	0.021	0.160	0.027

Table 10: The fairness scores of the inspected LMs on the task of hate speech detection. Teal color denotes the most fair model for each sensitive attribute according to each fairness metric (Elsafoury et al., 2023).

different fairness metrics give different fairness scores. However, there is a general trend that for the gender sensitive attribute, RoBERTa is the fairest, as for the race sensitive attribute, AlBERT is the fairest, and for the religion sensitive attribute, BERT is the fairest according to the majority of the fairness metrics.

To measure the impact of the SOS bias on the fairness scores, we measure the Pearson's correlation coefficient (ρ) between fairness scores measured by the different fairness metrics and the SOS_{LM} bias scores. Additionally, we measure the impact of the social bias in the LMs as measured by the CrowS-Pairs, StereoSet, and SEAT metric and the fairness of the LMs on the task of hate speech detection. The correlation results in Figure 4 show a consistent strong positive correlation between the CrowS-Pairs bias scores with the fairness scores measured by all three fairness metrics (FPR_gap. TPR_gap and AUC_gap) for all the models and sensitive attributes. And a less strong positive correlation with TPR_gap. There is a consistent negative correlation between SEAT scores and all fairness metrics. On the other hand, there is an inconsistent correlation with the StereoSet scores. As for the SOS bias, we find a positive correlation between the SOS_{LM} bias scores, against marginalized identities, and the fairness scores as measured by the FPR_gap and the AUC_gap. The results of this section suggest that the SOS bias in the LMs as measured by SOS_{LM} has an impact on the fairness of the LMs on the task of hate speech detection.

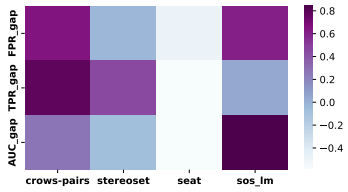

Figure 4: Heatmap of Pearson's correlation between social and SOS bias in LMs and fairness scores of LMs on the downstream task of hate speech detection.

8 Limitations

One of the main limitations of this work is that we only study bias in Western societies where Women, LGBTQ and Non-White ethnicities are among the marginalized groups. However, marginalized groups could include different groups of people in other societies. We also only use datasets, word lists, and LMs in English, which limits our study to the English-speaking world. Another limitation is that the correlation scores reported in the paper are not statistically significant, which could be due to having a few data points. Another limitation of this work is using a generative model to generate a list of profane and non-profane words. These generated word lists might be biased and might have an impact on the measured SOS bias scores. Moreover, the use of template sentence-pairs to measure the SOS bias in LMs do not provide a real context which might have impacted the measured SOS bias. The findings of this work are limited to the examined word embeddings, models, datasets, word lists, and might not generalize to others.

9 Conclusion

In this paper, we build on existing social bias metrics, and propose the SOS_{LM} metric to measure the systematic offensive stereotyping (SOS) bias in Language models (LMs) regarding six different sensitive attributes. Our results show that all the inspected LMs are SOS biased and that for the majority of the sensitive attributes, the SOS bias in the LMs is higher against the marginalized groups. Then, we validate the proposed SOS_{LM} metric in comparison to social bias metrics and published statistics on online hate that marginalized groups experience. Our results show that the proposed SOS_{LM} metric does not reveal different information from the social bias metric, especially CrowS-Pairs. But our results also show that the proposed metric to measure the SOS bias in LMs is reflective of the online hate experienced by marginalized

groups online. Subsequently, we use a state-of-the-art debias method, SentDebias, to remove the SOS bias. However, we found that SentDebias improved the SOS bias scores for some sensitive attributes and improved it for others. Finally, we investigate the impact of the SOS bias in LMs on their performance and fairness on hate speech detection. Our results suggest that there is an impact of the SOS bias in the LMs on their fairness on hate speech detection. However, there is no strong evidence that SOS bias has an impact on the performance of hate speech detection.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis.(2010). *Computational Statistics, John Wiley and Sons*, pages 433–459.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Fatma Elsafoury. 2022. Darkness can not drive out darkness: Investigating bias in hate speechdetection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 31–43. Association for Computational Linguistics.

Fatma Elsafoury, Stamos Katsigiannis, and Naeem Ramzan. 2023. On bias and fairness in nlp: How to have a fairer text classification? *arXiv preprint arXiv:2305.12829*.

Fatma Elsafoury, Stamos Katsigiannis, Steven R. Wilson, and Naeem Ramzan. 2021. Does BERT pay attention to cyberbullying? In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1900–1904, New York, NY, USA. Association for Computing Machinery.

Fatma Elsafoury, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: Systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1263–1274, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

James Hawdon, Atte Oksanen, and Pekka Räsänen. 2015. Online extremism and online hate. *Nordicom-Information*, 37:29–37.

Kaggle. 2012. Detecting insults in social commentary. `https://www.kaggle.com/c/detecting-insults-in-social-commentary/data`. Accessed: 2020-09-28.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 622–628. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399. International World Wide Web Conferences Steering Committee.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji rong Wen. 2023. Large language models for information retrieval: A survey. *ArXiv*, abs/2308.07107.

A Appendix

Attribute	Marginalized	Non-marginalized
Gender	woman, female, girl, wife, sister, daughter, mother	man, male, boy, son, father, husband, brother
Race	african, african american, asian, black, hispanic, latin, mexican, indian, middle eastern, arab	white, caucasian, european, american, european, norwegian, german, australian, english, french, american, swedish, canadian, dutch
Sexual-orientation	lesbian, gay, bisexual, transgender, tran, queer, lgbt,lgbtq,homosexual	hetrosexual, cisgender
Religion	jewish,buddhist,sikh, taoist, muslim	catholic, christian, protestant
Disability	blind, deaf, paralyzed	
Social-class	secretary, miner, worker, machinist, nurse, hairstylist, barber, janitor, farmer	writer, designer, actor, Officer, lawyer, artist, programmer, doctor, scientist, engineer, architect

Table 11: The non-offensive identity (NOI) words used to describe the marginalized and non-marginalized groups in each sensitive attribute. For the disability-sensitive attributes, we use only words to describe disability due to the lack of words used to describe able-bodied.