SOS: Systematic Offensive Stereotyping Bias in Word Embeddings

Fatma Elsafoury[1], Steven R. Wilson[2], Stamos Katsigiannis[3], and Naeem Ramzan[1]

[1]School of Physics, Engineering, and Computing, the University of The West of Scotland, UK
[2]Department of Computer Science and Engineering, Oakland University, USA
[3]Department of Computer Science, Durham University, UK

Abstract

Systematic Offensive stereotyping (SOS) in word embeddings could lead to associating marginalised groups with hate speech and profanity, which might lead to blocking and silencing those groups, especially on social media platforms. In this work, we introduce a quantitative measure of the SOS bias, validate it in the most commonly used word embeddings, and investigate if it explains the performance of different word embeddings on the task of hate speech detection. Results show that SOS bias exists in almost all examined word embeddings and that the proposed SOS bias metric correlates positively with the statistics of published surveys on online extremism. We also show that the proposed metric reveals distinct information compared to established social bias metrics. However, we do not find evidence that SOS bias explains the performance of hate speech detection models based on the different word embeddings.

1 Introduction

Wagner et al. (2021) describe *algorithmically infused societies* as the societies that are shaped by algorithmic and human behaviour. The data collected from these societies carry the same bias in algorithms and humans, like population bias and behavioural bias (Olteanu et al., 2019). These biases are important in the field of natural language processing (NLP) because unsupervised models like word embeddings encode them during training (Brunet et al., 2019; Joseph and Morgan, 2020). This includes racial bias which measures stereotypes related to people from different races, e.g. "Asians are good at math" (Garg et al., 2018; Manzini et al., 2019; Sweeney and Najafian, 2019), and gender bias which measures gender stereotypes, e.g. "women are housewives" (Garg et al., 2018; Bolukbasi et al., 2016; Chaloner and Maldonado, 2019). However, one aspect of bias that has received less attention is offensive stereotyping toward marginalised groups. For example, using slurs to describe non-white or LGBTQ communities or using swear words to describe women. Recent social research shows that using racial slurs and third-person profanity to describe groups of people aims at stressing the inferiority of the identity of the marginalised group (Kukla, 2018). Hence, as the internet is rife with slurs and profanity, it is important to study how machine learning models encode this offensive stereotyping.

To this end, we extend our initial work on introducing a computational measure of *systematic offensive stereotyping* (SOS) bias and examine its existence in pre-trained word embeddings (Elsafoury, 2022). We define SOS from a statistical perspective as "*A systematic association in the word embeddings between profanity and marginalised groups of people*". In other words, SOS refers to associating slurs and profane terms with different groups of people, especially marginalised people, based on their ethnicity, gender, or sexual orientation. Studies that focused on similar types of bias in hate speech detection models studied it within hate speech datasets themselves (Dixon et al., 2018; Waseem and Hovy, 2016a; Zhou et al., 2021), but not in the widely-used word embeddings which are, in contrast, not trained on data specifically curated to contain offensive content. Although some studies demonstrated that there is no correlation between intrinsic bias and extrinsic bias (Goldfarb-Tarrant et al., 2021), studying intrinsic bias on its own is an important task that reveals meaningful information about the data that was used to train those models, and in turn can help to expose harmful biases in society (Garg et al., 2018; Kambhatla et al., 2022).

In this work, we are interested in answering the following research questions: **RQ1:** How can we measure SOS bias? **RQ2:** What are the SOS bias scores of common pre-trained word embeddings,

Proceedings of the 29th International Conference on Computational Linguistics, pages 1263–1274
October 12–17, 2022.

and does SOS bias in the word embeddings differ from social biases? **RQ3:** How strongly does SOS bias correlate with external measures of online extremism and hate? **RQ4:** Does the SOS bias in the word embeddings explain the performance of these word embeddings on the task of hate speech detection? To answer our research questions, we build on the existing literature on measuring bias in word embeddings, propose a method to measure SOS bias, and investigate how different word embedding models associate profanity with marginalised groups.

Our contributions can be summarised as follows: (a) We define the SOS bias, propose a method to measure it in word embeddings and demonstrate that SOS bias correlates positively with the hate that marginalised people experience online. **(b)** We demonstrate that all the examined word embeddings contain SOS bias, with variations on the strength of the bias towards one particular marginalised group or another. **(c)** We show that there is no evidence that the SOS bias explains the performance of the different word embeddings on the task of hate speech detection. To allow more investigation on the topic, we share our code with the community [*].

2 Background

The term *bias* is defined and used in many different ways (Olteanu et al., 2019). There is the normative definition of bias, as its definition in cognitive science: "*behaving according to some cognitive priors and presumed realities that might not be true at all*" (Garrido-Muñoz et al., 2021). There is also the statistical definition of bias as "*systematic distortion in the sampled data that compromises its representatives*" (Olteanu et al., 2019).

In distributional word representations (Word Embeddings), the most common methods for quantifying bias are WEAT, RND, RNSB, and ECT: For WEAT, the authors were inspired by the Implicit Association Test (IAT) to develop a statistical test to demonstrate human-like biases in word embeddings (Caliskan et al., 2017). They used cosine similarity and statistical significance tests to measure the unfair correlations between two different demographic groups, as represented by manually curated word lists. For RND, the authors used the Euclidean distance between neutral words, like professions, and a representative group vector created by averaging the word vectors for words that describe a stereotyped group (gender/ethnicity) (Garg et al., 2018). In RNSB, a logistic regression model is first trained on the word vectors of unbiased labeled sentiment words (positive and negative) extracted from biased word embeddings. Then, that model was used to predict the sentiment of words that describe certain demographic groups (Sweeney and Najafian, 2019). In ECT, the authors proposed a method to measure how much bias has been removed from the word embeddings after debiasing (Dev and Phil, 2019).

These metrics, except RNSB, are based on the polarity between two opposing points, like male and female, allowing for binary comparisons. This forces practitioners to model gender as a spectrum between more "male" and "female" words, requiring an overly simplified view of the construct, leading to similar problems for other stereotypical types of bias, like racial, religious, transgender, and sexual orientation, where there are more than two categories that need to be represented (Sweeney and Najafian, 2019). These metrics also use lists of seed words that have been shown to be unreliable (Antoniak and Mimno, 2021). Since we are interested in measuring the systematic offensive stereotypes of different marginalised groups, these metrics would fall short of our needs. As for the RNSB metric, even though it is possible to include more than two identities, the sentiment dimension is represented as positive or negative (binary). But in our case, we are interested in a variety of offensive language targeted at different marginalised groups.

3 Systematic Offensive Stereotyping Bias

Our motivation is to reveal whether word embeddings associate offensive language with words describing marginalised groups. In the next section, we will use the SOS bias definition provided in the Introduction section to measure the SOS bias. For our experiments, we used 15 word embeddings: Word2Vec (W2V); Glove Wikipedia (Glove-WK); Glove-Twitter (Glove-Twitter); Urban Dictionary (UD); Chan word ; Glove Common Crawl (Glove-CC); Glove Common Crawl Large (Glove-CC-large); Fast-Text Common Crawl (FastText-CC); Fast-Text-Subwords Common Crawl (FT-CC-sws); Fast-Text Wiki (FT-Wiki); Fast-Text-Subwords wiki (FT-wiki-sws); sentiment specific word embedding (SSWE), Debias-W2V, P-DeSIP,

[*] https://github.com/efatmae/measure_SOS_bias_in_static_word_embeddings

Model	Dimensions	Trained on	Reference
W2V	300	100B words from Google News	(Mikolov et al., 2021a)
Glove-WK	200	6B tokens from Wikipedia 2014 and Gigaword	(Pennington et al., 2021)
Glove-Twitter	200	27B tokens collected from two billion Tweets	(Pennington et al., 2021)
UD	300	200M tokens collected from the Urban Dictionary website	(Urban dictionary, 2021)
Chan	150	30M messages from the 4chan and 8chan websites	(GSoC, 2019)
Glove-CC	300	42B tokens from Wikipedia 2014 and Gigaword	(Pennington et al., 2021)
Glove-CC-large	300	840B tokens from Wikipedia 2014 and Gigaword	(Pennington et al., 2021)
FastText-CC	300	600B common crawl tokens	(Mikolov et al., 2021b)
FT-CC-sws	300	600B common crawl tokens with subwords information	(Mikolov et al., 2021b)
FT-Wiki	300	16B tokens collected from Wikipedia 2017, UMBC, and statmt.org news dataset	(Mikolov et al., 2021b)
FT-wiki-sws	300	16 billion tokens with subwords information collected from the Wikipedia 2017, UMBC, and statmt.org	(Mikolov et al., 2021b)
SSWE	50	10M comments collected from Twitter	(Tang et al., 2014)
Debias-W2V	300	W2V model after the gender bias has been removed using the hard debiasing method	(Bolukbasi et al., 2016)
P-DeSIP	300	Debiased Glove-WK with the potential proxy gender bias removed.	(Ding et al., 2022)
U-DeSIP	300	Debiased Glove-WK word embeddings with the unresolved gender bias removed.	(Ding et al., 2022)

Table 1: Description of the word embeddings used in this work.

Group	Words
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

Table 2: Non-offensive identity (NOI) words and the group they describe.

and U-DeSIP. Table 1 provides information of the different word embeddings.

3.1 Measuring SOS bias

Based on our definition of SOS, to answer RQ1, we propose to measure the SOS bias using the cosine similarity between swear words and words that describe marginalised social groups. For the swear words, we used a list (Swear words, 2022) that contains 403 offensive expressions, reduced to 279 after removing multi-word expressions[†]. We used a non-offensive identity (NOI) word list to describe marginalised groups of people (Zhou et al., 2021; Dixon et al., 2018) and non-marginalised ones (Sweeney and Najafian, 2019), as summarised in Table 2. Unlike WEAT, ECT, and RND, which used seed words like people's names to infer their nationality or pronouns, we used NOI words to describe the different groups similar to the RNSB metric. According to (Antoniak and Mimno, 2021), using NOI words is a better motivated and more coherent approach for describing groups of people than names.

Let $W_{NOI} = \{w_1, w_2, w_3, ...w_n\}$ be the list of NOI words w_i, $i = 1, 2, ..., n$, and $W_{sw} = \{o_1, o_2, o_3, ...o_m\}$ be the list of swear words o_j, $j = 1, 2, ..., m$. For measuring the SOS bias for a specific word embedding we, firstly, we compute the average vector $\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}$ of the swear words for we, e.g. for W2V, etc. $SOS_{i,we}$ for a NOI word w_i and a word embedding we is then defined (Equation 1) as the cosine similarity between $\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}$ and the word vector $\overrightarrow{w_{i,we}}$, for the word embedding we, normalised to the range $[0, 1]$ using min-max normalisation across all NOI words (W_{NOI}), in order to ease comparison between the different word embeddings.

$$SOS_{i,we} = \frac{\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}} \cdot \overrightarrow{w_{i,we}}}{||\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}|| \cdot ||\overrightarrow{w_{i,we}}||} \quad (1)$$

The normalised SOS scores are in the range $[0, 1]$ and indicates the similarity of a NOI word to the average representation of swear words. Accordingly, a higher $SOS_{i,we}$ value for word w_i indicates that the word embedding $\overrightarrow{w_{i,we}}$ for the word w_i, is more associated with profanity. We intended for the metric to be used in a comparative manner among word embeddings, e.g. W2V vs Glove-WK, or among different groups of people, e.g. LGBTQ vs Straight, rather than to determine an objective threshold below which no bias exists.

We computed the mean SOS score for our examined word embeddings using the aforementioned swear words and NOI word lists for each examined group individually, as well as for the combined marginalised (Women, LGBTQ, Non-white

[†]We repeated the same experiment with a different set of 427 swear words from (Agrawal and Awekar, 2018) and also observed a significantly higher SOS bias scores for marginalised groups for 11 word embeddings.

Word embeddings	Mean SOS							
	Gender		Sexual orientation		Ethnicity		Marginalised vs. Non-marginalised	
	Women	Men	LGBTQ	Straight	Non-white	White	Marginalised	Non-marginalised
W2V	**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 3: 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).

ethnicities) and non-marginalised (Men, Straight, White ethnicities) groups. Table 3 shows that most of the word embeddings are more biased towards the marginalised groups than the non-marginalised groups, with some word embeddings being more SOS biased than others. It also shows that mean SOS bias scores towards the marginalised groups for all the word embeddings, except for Fast-text-wiki-subwords, are higher towards the non-marginalised groups (Wilcoxon $p = 0.0001, \alpha = 0.05$). For Fast-text-wiki-subwords, the SOS bias score for the non-marginalised groups (0.635) is marginally higher than the SOS bias score for the marginalised groups (0.632). In addition, the debiased word embeddings where gender information is removed (Debiased W2V, P-DeSIP, and U-DeSIP), still contain slightly higher SOS bias towards women than men. Given that SOS bias is significantly higher for marginalised groups (Table 3) and that most hate speech datasets contain hate towards women and the marginalised groups, this work subsequently focuses on those groups (Women, LGBTQ, Non-white).

3.2 SOS biased word embeddings

To answer the first part of RQ2, we conducted a comparative analysis of the word embeddings with regard to SOS bias. Table 4 shows the bias scores of each of the word embeddings towards each marginalised group. To quantitatively compare the different word embeddings, we used the SOS bias scores for each marginalised group (LGBTQ, Women, Non-white ethnicities) and applied different significance tests at $\alpha = 0.05$. The results in Table 4 show that Glove-twitter, Chan, Glove-CC, and Fast-text-wiki-subwords are the most biased towards women, with Chan being the most biased ($SOS_{women,Chan} = 0.88$), and Debias-W2V the least biased ($SOS_{women,Debias\text{-}W2V} = 0.205$), which could be due to the fact that Debias-W2V is W2V after removing gender bias. When we used the Friedman test to compare the SOS scores of the different word embeddings for the individual words that describe the "Women" group, the results showed a significant difference between the different word embeddings ($p = 2e^{-11}$), indicating that Chan is significantly more biased towards "Women" in comparison to the rest of the word embeddings. It is worth noting that the reduction in SOS_{women} from 0.435 for Glove-WK to 0.266 for P-DeSIP and U-DeSIP is higher than the reduction achieved for W2V (to Debias-W2V) from 0.293 to 0.205, meaning that U-DeSIP and P-DeSIP used more effective debiasing methods for this category. On the other hand, U-DeSIP and P-DeSIP have higher SOS bias scores toward non-white ethnicities than Glove-WK (as did Debias-W2V compared to W2V), indicating that while bias reduction methods decrease biases toward some groups, they may unintentionally *increase* bias towards others.

Word embeddings	Mean SOS		
	Women	LGBTQ	Non-white
W2V	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 4: 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.

The LGBTQ community is the group that is most biased against by most of the word embeddings, i.e. W2V, Glove-WK, UD, Fast-text-CC, Fast-text-wiki, P-DeSIP, and U-DeSIP. Glove-WK is the most biased ($SOS_{lgbtq,Glove\text{-}WK} = 0.669$), whereas the least biased is SSWE ($SOS_{lgbtq,SSWE} = 0.438$). When we used the Friedman test to compare the SOS scores of the different word embeddings for the individual words that describe the "LGBTQ" group, the results showed a significant difference between the different word embeddings ($p = 0.048$), indicating that Glove-WK is significantly more SOS biased towards the "LGBTQ" community in comparison to the other word embeddings. These findings are notable as Glove-WK was pre-trained on Wikipedia articles which are expected to have the least profanity compared to social media or common crawl.

Table 4 also shows that Glove-CC-large, Fast-text-CC-subwords, SSWE, and Debias-W2V are the most biased towards non-white ethnicities, with SSWE being the most biased ($SOS_{non\text{-}white,SSWE} = 0.688$) and Glove-WK the least biased ($SOS_{non\text{-}white,Glove\text{-}WK} = 0.234$). When we used the Friedman test to compare the SOS scores of the different word embeddings for the individual words that describe the "Non-white-ethnicities" group, the results showed a significant difference between the different word embeddings ($p = 3e^{-6}$), indicating that SSWE is significantly more biased towards "Non-white-ethnicities" in comparison to the rest of the word embeddings. Since SSWE was pre-trained on sentiment information, and as Sweeney and Najafian (2019) showed, the sentiment towards non-white ethnicities is mostly negative, our results are in line with earlier findings.

3.3 SOS bias and other social biases

In this section, we answer the second part of RQ2 by comparing our SOS bias scores to gender and racial bias as measured by existing social bias metrics from the literature (WEAT, RND, RNSB, ECT). We used the WEFE framework (Badilla et al., 2020) to measure the gender bias using the other state-of-the-art metrics and two target lists: Target list 1, which contained female-related words (e.g., she, woman, and mother), and Target list 2, which contained male-related words (e.g., he, father, and son), as well as two attribute lists: Attribute list 1, which contained words related to family, arts, appearance, sensitivity, stereotypical female roles, and negative words, and Attribute list 2, which contained words related to career, science, math, intelligence, stereotypical male roles, and positive words (Badilla et al., 2020; Caliskan et al., 2017). Then, we measured the average gender bias scores across the different attribute lists for each word embedding using the various metrics. For the SOS bias, we used the mean SOS scores of the words that belong to the "Women" category. Contrary to all the metrics, ECT scores have an inverse relationship with the level of bias, so we subtract all ECT scores from 1 to enforce that higher scores for all metrics indicate greater levels of bias. We then computed the Spearman's rank correlation coefficient between the gender bias scores of the different word embeddings, as measured by WEAT, RND, RNSB, ECT, SOS_{women}.

To measure the racial bias using the state-of-the-art metrics, we used two target groups: Target group 1, which contained stereotypical white names, and Target group 2, which contained stereotypical African, Hispanic, and Asian names, and two attribute lists: Attribute list 1, which contained white people's occupation names; and Attribute list 2, which contained African, Hispanic, and Asian people's occupations (Badilla et al., 2020; Garg et al., 2018). Then, we measured the average racial bias scores across the different attribute lists for each word embedding using the different metrics (WEAT, RND, RNSB, ECT). For the SOS bias, we used the mean SOS scores of the words that belong

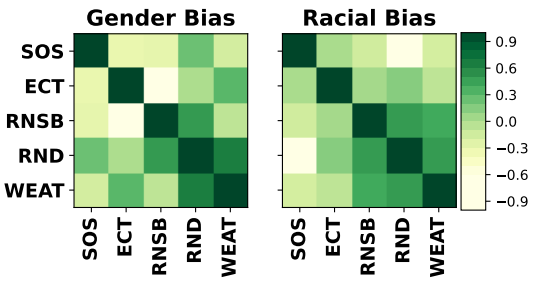

Figure 1: Spearman's correlation between the different bias metrics (SOS and social bias) for all the examined word embeddings. For gender bias, SOS refers to SOS_{women}, and for racial bias to $SOS_{non\text{-}white}$.

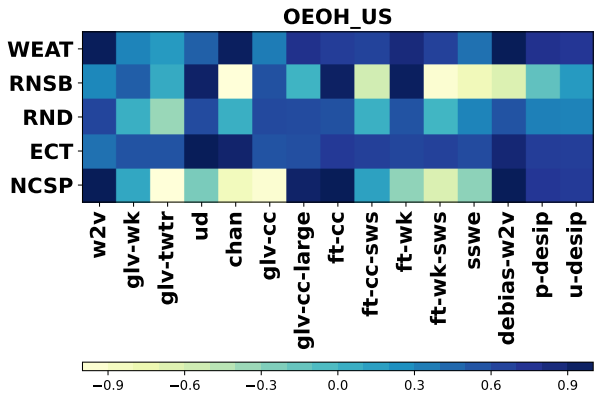

Figure 2: Pearson's correlation between the different SOS bias metrics and the percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online for the OEOH-US survey for the word embeddings.

to the "Non-white ethnicities" category. Finally, we computed the Spearman's rank correlation coefficient between the different racial bias scores of the different word embeddings, as measured by WEAT, RND, RNSB, ECT, $SOS_{non\text{-}white}$.

The results in Figure 1 show that for gender bias, WEAT has a strong positive correlation with RND and a positive correlation with ECT and RNSB. On the other hand, SOS has almost no correlation with ECT, RNSB, WEAT and a small positive correlation with RND. For racial bias, WEAT has a positive correlation with RNSB, and RND, no correlation with ECT and a negative correlation with SOS. On the other hand, SOS has a negative correlation with RNSB, RND, and WEAT and almost no correlation with ECT. The results here suggest that the SOS bias reveals different information than the social bias metrics, especially for racial bias. We speculate that this is the case because profanity is more often used online with non-white ethnicities than with women (Hawdon et al., 2015).

3.4 SOS bias validation

To answer RQ3, we compared the SOS bias measured by our proposed method, as well as by existing metrics (WEAT, RNSB, RND, ECT), to published statistics on online hate and extremism that is targeted at marginalised groups (Women, LGBTQ, Non-white ethnicities). To avoid confusion since all metrics measure SOS bias in this case, we refer to our proposed method for measuring SOS bias as "*normalised cosine similarity to profanity*" or **NCSP** for short. We used the WEFE framework (Badilla et al., 2020) to measure the SOS bias of the examined word embeddings using the state-of-the-art metrics. The metrics in the WEFE platform take 4 inputs: Target list 1: a word list describing a group of people, e.g. women; Target list 2: a word list that describes a different group of people, e.g. men; Attribute list 1: a word list that contains attributes that are believed to be associated with target group 1, e.g. housewife; and Attribute list 2: a word list that contains attributes that are believed to be associated with target group 2, e.g. engineer. Each metric then measures these associations, as described in Section 2.

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 5: The percentage of examined groups that experience online hate and extremism in different countries (Hawdon et al., 2015)

To measure the SOS bias for gender using the state-of-the-art metrics, target list W1 contained the NOI words that describe women from Table 2, target list W2 contained the NOI words that describe men, attribute list 1 contained the same swear words used earlier to measure our SOS bias (Section 3.1), and attribute list 2 a list of positive words provided by the WEFE framework. To measure the SOS bias for ethnicity using the state-of-the-art metrics, we used the same process, with the same attribute lists, but with target list E1 that contained NOI words that describe non-white ethnicities and target list E2 that contained NOI words that describe white ethnicities. Similarly, to measure the SOS bias for sexual orientation, we used the same attribute lists and target list L1, which contained NOI words that describe LGBTQ people, and target list L2 which contained NOI words that describe straight people. To measure the SOS

Dataset	Samples	Positive samples
HateEval	12722	42%
Twitter-sexism	14742	23%
Twitter-racism	13349	15%
Twitter-hate	5569	25%

Note: Positive samples refer to offensive comments

Table 6: Hate speech datasets' details.

bias for gender, ethnicity, and sexual orientation with our proposed metric (NCSP), we computed the mean SOS scores of the NOI words that describe women, LGBTQ, and non-white for each word embeddings as in Table 4.

The percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online were then acquired from the online extremism and online hate survey (OEOH), collected by (Hawdon et al., 2015) from Finland, Germany, the US, and the UK in 2013 and 2014, for individuals aged 15-30. Table 5 provides details on the published statistics. Then, we computed the Pearson's correlation coefficient between the SOS[‡] scores, measured by the different metrics for Women, LGTBQ, and Non-white ethnicities for the examined word embeddings and the percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online. Figure 2[§] shows that the SOS bias correlates positively with the published statistics on online hate and extremism.

When we first look at the different metrics for measuring the SOS bias, we find that bias metrics like WEAT, RND, and ECT correlate more positively with the OEOH survey in the US. However, when we look closely at the order of the percentages of marginalised groups regarding their experience of online hate, we find that the LGBTQ community experiences online hate the most, followed by non-white ethnicities with a marginal difference, and then women. Consequently, we expect that the survey results would correlate strongly positively with the word embeddings that are least biased towards women (e.g. W2V, FT-CC, Debias-W2V, P-DeSIP, and U-DeSIP); correlate less positively with word embeddings that are more biased towards women than LGBTQ or Non-white (e.g. Glove-WK, UD, FT-WK, and SSWE); and correlate negatively with word embeddings that are most biased towards women (e.g. Glove-twitter, Chan, Glove-CC, FT-WK-sws).

This pattern of correlation is achieved only by our proposed metric, which reflects the variation of the SOS bias scores towards the different marginalised groups in each word embedding, in comparison to WEAT, ECT and RND, which do not reflect these variations and hence correlate indiscriminately positively with all the word embeddings. RNSB does reflect some of that variation but not as consistently as our proposed metric. The results suggest that our proposed metric for measuring SOS bias (NCSP) is the most reflective of the SOS bias in the different word embeddings.

4 SOS bias and hate speech detection

In this section, we answer RQ4 through a series of experiments on hate speech detection. We trained deep learning models with an embedding layer for the detection of hate speech from hate speech-related datasets, then computed the correlation of the performance of the different word embeddings to the SOS bias score of these embeddings. We used four hate-speech-related datasets that contain different types of hate speech (Table 6): (i) *Twitter-racism*, a collection of tweets labeled as racist or not (Waseem and Hovy, 2016b); (ii) *Twitter-sexism*, tweets labeled as sexist or not (Waseem and Hovy, 2016b); (iii) Twitter-hate, containing tweets labeled as offensive, hateful (sexist, homophobic, and racist), or neither (Davidson et al., 2017), but as we are interested in the hateful content, we used the tweets that are labeled as hateful or neither; and (iv) *HateEval*, a collection of tweets containing hate against immigrants and women in Spanish and English (Basile et al., 2019), from which we used only the English tweets. These four datasets were selected because they contain hate speech towards the marginalised groups that are the focus of our study thus they are representative of the examined problem.

To pre-process the datasets, we removed URLs, user mentions, retweet abbreviation "RT", non-ASCII characters, and English stop words except for second-person pronouns like "you/yours/your", and third-person pronouns like "he/she/they", "his/her/their" and "him/her/them", as suggested in (Elsafoury et al., 2021). All letters were lowercased, and common contractions were converted to their full forms. And each dataset was randomly

[‡]We subtract all ECT scores from 1 here as well.

[§]The correlation results for OEOH-US are similar to the correlation results from OEOH-Finland, OEOH-UK and OEOH-Germany, and thus are omitted from Figure 2.

Word embeddings	HateEval		Twitter-Hate		Twitter-racism		Twitter-sexism	
	MLP	BiLSTM	MLP	BiLSTM	MLP	BiLSTM	MLP	BiLSTM
W2V	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 7: F1 scores for the used models for hate speech detection using the examined word embeddings on the examined datasets. Bold values indicate the highest scores among the different word embeddings per model and dataset.

Dataset	Model	WEAT	RNSB	RND	ECT	NCSP
HateEval	MLP	**0.277**	0.223	-0.100	0.019	0.230
	BiLSTM	0.377	**0.540***	0.094	-0.030	0.100
Twitter Sexism	MLP	**0.157**	0.030	-0.216	-0.039	0.121
	BiLSTM	0.109	**0.266**	0.093	-0.361	0.246
Twitter Racism	MLP	0.042	0.017	-0.336	-0.223	**0.241**
	BiLSTM	-0.264	**0.135**	-0.210	-0.103	0.110
Twitter Hate	MLP	0.107	0.218	-0.164	-0.148	**0.223**
	BiLSTM	**0.507**	0.475	0.289	-0.217	0.396

Table 8: 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. * indicates that the correlation is statistically significant at $p < 0.05$.

split into a training (70%) and a test (30%) set, preserving class ratios.

We used two deep learning models: (i) a Bidirectional LSTM (Schuster and Paliwal, 1997) with the same architecture as in (Agrawal and Awekar, 2018), who used RNN models to detect hate speech, and (ii) a two-layer Multi-Layer Perceptron (MLP) model. To this end, we first used the Keras tokenizer (Tensorflow.org, 2020) to tokenise the input texts, using a maximum input length of 64 (maximum observed sequence length in the dataset). A frozen embedding layer, based on a given pretrained word embedding model, was used as the first layer and fed to the BiLSTM model and the MLP model. To avoid over-fitting, we used L2 regularisation with an experimentally determined value of 10^{-7}. The models were trained for 100 epochs with a batch size of 32, using the Adam optimiser and a learning rate of 0.01 (default of Keras Optimiser) (Agrawal and Awekar, 2018). For each dataset, we used a 5-fold cross-validation to train and validate a model (70% and 30% of the training set respectively, with class ratio preserved) and then test each fold's model on the test set. Then, the average F1-score across the 5 folds was reported.

4.1 Experimental Results

Given the results for the SOS bias in the different embeddings (Table 4), we hypothesise that the deep learning models that are trained with Glove-CC-large, FastText-CC-subwords, SSWE, and Debias-W2V embeddings will perform the best (highest F1 score) on datasets that contain hate speech or insults towards marginalised ethnicities, which is Twitter-racism. We also hypothesise that the models trained with Glove-Twitter, Chan, Glove-CC, and Fast-text-wiki-subwords will achieve the highest F1 scores on datasets that contain insults towards women, which is Twitter-sexism. Since Twitter-Hate and HateEval contain a mixture of hateful content towards women and immigrants, we hypothesise that the best performing word embeddings would be the ones that have SOS scores higher than the median values for both of SOS_{women}

(0.473) and $SOS_{Non\text{-}white}$ (0.456), which are Glove-Twitter, Fast-text-wiki-subwords, and SSWE.

The performance of the deep learning models with the different embedding models is reported in Table 7. The results show that for all datasets, BiLSTM outperforms MLP in terms of F1 score. The results also show that for the MLP model, our hypotheses hold for the Twitter-racism dataset, as the best performing models are BiLSTM with Fast-text-CC-subwords and MLP with Glove-CC-large. However, for Twitter-sexism, HateEval, and Twitter-Hate, the results do not support our hypothesis, with Fast-text-CC and Glove-CC-large being the best performing with MLP and BiLSTM models. To quantify our analysis we used the Spearman's correlation between the SOS bias scores, measured using the different bias metrics, of the different word embeddings and the F1 scores of the MLP and BiLSTM trained with the different word embeddings. The results in Table 8 show occasionally positive correlations for example with WEAT, RNSB, and our proposed metric NCSP. However, most of these positive correlations are not statistically significant except for the SOS scores measured by the RNSB metric and the F1 of the BiLSTM model and the HateEval dataset. These results show that there is no positive correlation between the SOS bias scores in the word embeddings and the performance of the hate speech detection models, suggesting that the SOS bias in the word embeddings does not explain their utility as features for hate speech detection.

5 Conclusion

In this work, we built on our initial work (Elsafoury, 2022) where the SOS bias was introduced and proposed methods to measure it, validate it, compare it to stereotypical social bias, and investigate if it explains the performance of the word embeddings on hate speech detection. Results show that the examined word embeddings are SOS biased and that the SOS bias in the word embeddings has a strong positive correlation with published statistics on online extremism. However, more datasets need to be collected to provide stronger evidence, especially data from social sciences on the offences that marginalised groups receive on social media. Nonetheless, this is an informative finding as it reveals the bias in the dataset that these word embeddings were trained on. Since not all these datasets are available to the public, measuring the SOS bias in the word embeddings is an important way to learn about that bias in those datasets.

Our findings also show that the proposed SOS bias reveals different information than the types of bias measured by existing metrics. Finally, our findings show no evidence that the SOS bias, measured using different bias metrics, explains the performance of the different word embeddings on the task of hate speech detection. This finding suggests that the SOS bias, and potentially other biases in general, are not strongly related to word embeddings' performance on the downstream task of hate speech detection. We plan to examine this speculation and study the influence of the SOS and social bias on the *fairness* of hate speech detection models in future work.

6 Limitations

The findings demonstrated in this paper are limited to the inspected word embeddings, models, and datasets, and might not generalise to other datasets. Similarly, our SOS bias scores are limited to the used word lists and even if we used two different swear word lists and identity terms that are coherent according to (Antoniak and Mimno, 2021), using different word lists may give different results. Another limitation is regarding our definition of the SOS bias, as we defined bias from a statistical perspective which lacks the social science perspective as discussed in (Blodgett et al., 2021; Delobelle et al., 2022). Moreover, we only studied bias in Western societies where Women, LGBTQ and Non-White ethnicities are among the marginalised groups. However marginalised groups could include different groups of people in other societies. We also only used datasets and word lists in English which limits our study to the English speaking world. Similar to other works on quantifying bias, our proposed metric measures the existence of bias and not its absence (May et al., 2019), and thus low bias scores do not necessarily mean the absence of bias or discrimination in the word embeddings.

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 141–153. Springer.

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.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. WEFE: the word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 430–436. ijcai.org.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna M. Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1004–1015. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard S. Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR.

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

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1693–1706. Association for Computational Linguistics.

Sunipa Dev and Jeff M. Phil. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR.

Lei Ding, Dengdeng Yu, Jinhan Xie, Wenxing Guo, Shenggang Hu, Meichen Liu, Linglong Kong, Hongsheng Dai, Yanchun Bao, and Bei Jiang. 2022. Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11864–11872. AAAI Press.

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, 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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings*

of the National Academy of Sciences, 115(16):E3635–E3644.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7).

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.

GSoC. 2019. 4 and 8 chan embeddings. [Online] Accessed 05/11/2021.

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

Kenneth Joseph and Jonathan Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online. Association for Computational Linguistics.

Gauri Kambhatla, Ian Stewart, and Rada Mihalcea. 2022. Surfacing racial stereotypes through identity portrayal. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1604–1615. ACM.

Rebecca Kukla. 2018. Slurs, interpellation, and ideology. *The Southern Journal of Philosophy*, 56:7–32.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2021a. word2vec embeddings. [Online] Accessed 05/11/2021.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2021b. Glove twitter embeddings. [Online] Accessed 25/04/2022.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2021. Glove twitter embeddings. [Online] Accessed 05/11/2021.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Swear words. 2022. Swear words list. [Online] Accessed 26/04/2022.

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.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Tensorflow.org. 2020. Text tokenization utility class. `https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer`. Accessed: 2020-09-28.

Urban dictionary. 2021. Urban dictionary embeddings. [Online] Accessed 05/11/2021.

Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. 2021. Measuring algorithmically infused societies. *Nature*, 595(7866):197–204.

Zeerak Waseem and Dirk Hovy. 2016a. 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.

Zeerak Waseem and Dirk Hovy. 2016b. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A. Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3143–3155. Association for Computational Linguistics.