SOS: Systematic Offensive Stereotyping Bias in Word Embeddings

Anonymous ACL submission

Abstract

Hate speech detection models aim to provide a safe environment for marginalised social groups to express themselves. However, the bias in these models could lead to silencing those groups. In this paper, we introduce the systematic offensive stereotyping (SOS) bias. We propose a method to measure the SOS bias in different word embeddings and also investigate its influence on the downstream task of hate speech detection. Our results show that SOS bias against various groups exists in widely used word embeddings and that our SOS bias metric correlates positively with the statistics of published surveys on online abuse and extremism. However, we found that it is not easy to prove that bias in word embeddings influences downstream task performance. Finally, we show that SOS bias is more indicative of sexism and racism in the inspected word embeddings when used for sexism and racism detection than social biases.

1 Introduction

Wagner et al. (2021) describe the term *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 (Garg et al., 2018; Manzini et al., 2019; Sweeney and Najafian, 2019), gender bias (Garg et al., 2018; Bolukbasi et al., 2016; Chaloner and Maldonado, 2019), and personality stereotypes (Agarwal et al., 2019). However, one aspect of bias that has received less attention is systematic offensive stereotyping (SOS) in word embeddings. 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 offensive terms to 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 (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. Moreover, most studies on bias in word embeddings focused on Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). However, recent word embeddings models, like the Urban Dictionary word embeddings, pre-trained on words and definitions from the Urban Dictionary website (Wilson et al., 2020), the Chan word embeddings, pre-trained on the 4&8 Chan websites (Voué et al., 2020), and a version of GloVe pre-trained on Twitter data (Stojanovski et al., 2015), have received much less attention in previous studies of bias. The social media platforms on which these embeddings have been trained are biased (Nguyen et al., 2017; Voué et al., 2020; Mittos et al., 2020; Mislove et al., 2011). Additionally, the literature on bias in word embeddings claims that it influences downstream tasks, like translation, text classification, and text generation. Still, these claims have not yet been tested (Blodgett et al., 2020).

In this work, we are interested in answering the following research questions: **RQ1:** How can we measure the SOS bias? **RQ2:** Among the examined word embedding models, which has the most SOS bias? **RQ3:** How strongly does SOS bias correlate with external measures of online extremism and abuse? **RQ4:** How does SOS bias in word embeddings relate to performance on downstream tasks? **RQ5:** How does SOS bias differ from stereotypical social bias regarding finding the most biased word embeddings when used for the task of hate speech detection? To answer our research questions, we built on the existing literature on measuring bias in word embeddings and proposed a method to measure SOS bias and investigate how different word embedding models associate profanity with marginalised people. **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 our SOS bias is representative of the abuse that marginalised people experience online. (b) We demonstrate that all the examined word embeddings are SOS biased, with variations on the strength of the bias towards one particular marginalised group or another. (c) We demonstrate that the claim that bias in word embeddings influences downstream tasks is not easy to prove and that despite finding a positive correlation between the SOS bias scores and the performance on the downstream tasks, it is not conclusive. (d) We demonstrate that SOS bias is more indicative of the sexism and racism in the inspected word embeddings than the stereotypical social bias, gender, and racial biases, as measured by state-of-the-art metrics when used for the task of hate speech detection. (e) We share our code with the community.

Our findings show that the different word embeddings are SOS biased, particularly towards marginalised groups, and it does have an influence, to some extent, on the downstream tasks of hate speech and abuse detection. This bias could have negative implications as these hate speech detection models might learn to associate marginalised groups with extremism and abuse. As a result, these models that were supposed to provide a protective environment for the marginalised people to express themselves are the ones that could lead to silencing them or flagging their content as inappropriate.

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 as: "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 the case of bias 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 the cosine similarity and statistical significance tests to measure the unfair correlations for two different demographics, 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 has 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 demographics (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 them (Dev and Phillips, 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 and to answer RQ1. For our experiments, we used five word embeddings: Word2vec (w2v), trained on a collection of 100 billion words from Google News (Mikolov et al., 2021); Glove Wikipedia (Glove-WK), trained on a collection of six billion tokens from Wikipedia 2014 and Gigaword (Pennington et al., 2021b); Glove-Twitter (Glove-Twitter), trained on 27 billion tokens collected from two billion Tweets (Pennington et al., 2021a); Urban Dictionary (UD), trained on 200 million token collected from the Urban Dictionary website (Urban dictionary, 2021); and Chan word embeddings, trained on 30 million messages from the 4chan and 8chan websites (GSoC, 2019).

3.1 Measuring SOS bias

Based on our definition of SOS, we want a method to measure the association that each word embedding model has between profanity and marginalised groups of people. To answer RQ1, we propose to measure that association using the cosine similarity between swear words and words that describe marginalised social groups. For the swear words,

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

*Marginalised group

Table 1: NOI words and the group they describe.

we used a list of 427 swear words from (Agrawal and Awekar, 2018; Dinakar et al., 2011). For describing marginalised social groups, we used a word list that contains non-offensive identity (NOI) names 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 1. Similar to RNSB, we use NOI words to describe the different groups, unlike WEAT, ECT, and RND which used seed words like people's names to infer their nationality or pronouns. The motivation behind using NOI words is clearer than using seed words used in the literature (Antoniak and Mimno, 2021). And even though

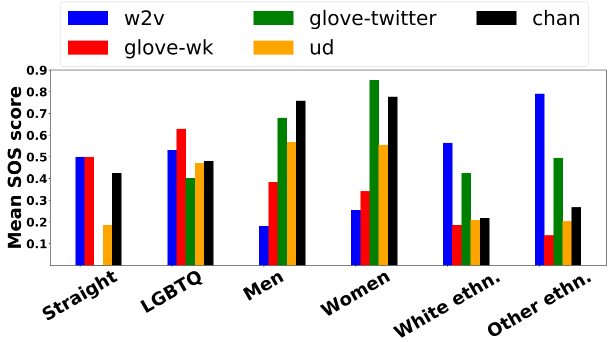

Figure 1: Mean SOS scores for the examined word embeddings and groups.

our NOI words that describe the same groups e.g. Non-white ethnicities have not been examined for coherence using semantic similarity for example as suggested by (Antoniak and Mimno, 2021), our NOI words' groups are more coherent than the seed words used in the literature which used people's names to describe African Americans or Asian nationalities.

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, $j = 1, 2, ..., m$. To measure the SOS bias for a specific word embedding we, we first compute the average vector $\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}$ of the swear words for we, e.g. for Word2Vec, Glove, 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}).

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

The normalised SOS score takes values within the range $[0, 1]$ and indicates the similarity of a NOI word to the average representation of swear words. Consequently, 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. The metric is intended to be used in a comparative manner among word embeddings, e.g. w2v vs Glove-WK, or among different groups of people, e.g. Women vs Men, rather than to determine an objective threshold below which no bias exists.

3.2 Mean SOS for word embeddings

We computed the mean SOS score for our examined word embeddings(Word2Vec, Glove-WK, Glove-Twitter, UD, and Chan) 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 ethnicities) and non-marginalised (Men, Straight, White ethnicities) groups. Figure 1 shows that some word embeddings are more biased than others and that the biased word embeddings are more biased towards the marginalised group than the non-marginalised groups. In addition, Table 2 shows that mean SOS bias towards the marginalised groups is higher than towards the non-marginalised groups (T-test $p = 0.02, \alpha = 0.05$).

It is also evident that when comparing the "Straight" and the "LGBTQ" groups, there is a higher SOS bias towards the marginalised "LGBTQ" group for all the examined word embeddings. Similar for the "Men" vs. "Women" groups and "White ethnicity" vs. "Other ethnicities" groups, where there is higher SOS bias towards the marginalised "Women" and "Other ethnicities" groups, except for Glove-WK and UD for which the SOS bias is marginally higher for the non-marginalised groups ("Men", "White ethnicity"). Given that SOS bias is significantly higher for marginalised groups (Table 2) and most of the hate speech datasets contain hate towards the marginalised groups, this work subsequently focuses on those groups (women, lgtbq, non-white).

Word embedding	Mean SOS	
	Marginalised	Non-marginalised
Word2Vec	**0.535**	0.430
Glove-WK	**0.390**	0.281
Glove-Twitter	**0.558**	0.469
UD	**0.407**	0.325
Chan	**0.495**	0.417

Table 2: Mean SOS score of the different groups.

3.3 SOS biased word embeddings

To answer RQ2, we conducted a comparative analysis between the word embeddings in regards to SOS bias. To quantitatively compare the different word embeddings, we used the SOS bias scores (Figure 1) for each marginalised group (LGTBQ, Women, Other ethnicities) and applied the Friedman and T-test significance tests ($\alpha = 0.05$). For the words that describe the "LGTBQ" group, Glove-WK has the highest SOS score of 0.629, but the Friedman test failed in finding a significant difference between the different word embeddings ($p = 0.6$), indicating that all the examined word embeddings are similarly SOS-biased towards words related to the "LGBTQ" group. For the "Women" group, Glove-Twitter, UD, and Chan exhibited high SOS bias, with Glove-Twitter having the highest score of 0.852, and Friedman's test indicating a significant difference between the word embeddings ($p = 5e^{-4}$). A T-test showed that Glove-Twitter is significantly different from Word2Vec, Glove-WK, and UD ($p = 6e^{-6}, 1e{-}5,$ and 0.0057 respectively), but no significant difference from Chan ($p = 0.350$) could be established. This indicates that Glove-Twitter and Chan exhibit a similar significant SOS bias towards women (sexism) in comparison to Word2Vec, Glove-WK, and UD. Regarding the "Other ethnicities" group, Word2Vec stands out as the word embedding with the highest SOS score of 0.691. Friedman's test showed a statistically significant difference between all the word embeddings ($p = 4e{-}4$) and the T-test showed that the SOS score of Word2Vec is significantly higher than Glove-WK, Glove-Twitter, UD, and Chan ($p = 9e^{-7}, 8e^{-3}, 1e^{-5},$ and $4e^{-5}$ respectively), indicating that Word2Vec is significantly SOS-biased towards non-white ethnicities in comparison to Glove-WK, Glove-Twitter, UD, and Chan. We summarise our results in Table 3 showing that Word2Vec is the most SOS-biased towards non-white ethnicities, Glove-WK is the most SOS-biased towards the LGBTQ community, and Glove-Twitter, UD, and Chan are the most SOS-biased towards women.

Word Embedding	SOS biased towards
Word2Vec	**Other ethnicities**, LGBTQ, Women
Glove-WK	**LGBTQ**, Women, Other ethnicities
Glove-Twitter	**Women**, Other ethnicities, LGBTQ
UD	**Women**, LGBTQ, Other ethnicities
Chan	**Women**, LGBTQ, Other ethnicities

Table 3: The groups that each word embedding is SOS-biased towards, ordered by descending severity.

3.4 SOS bias validation

To answer RQ3, we compared the SOS bias, measured by our proposed method and state-of-the-art metrics (WEAT, RNSB, RND, ECT), to published statistics on online abuse and extremism that is targeted at marginalised groups (Women, LGBTQ, Non-white ethnicities). The WEFE framework (Badilla et al., 2020) was used 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

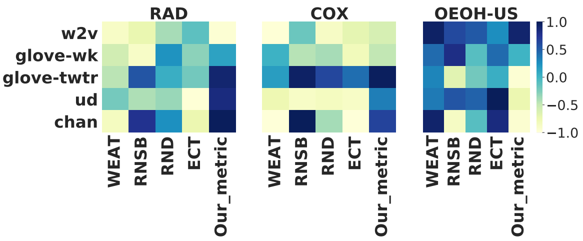

Figure 2: The Pearson's correlation between the different metrics and the percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online for each published survey for the examined word embedding. For RAD heatmap, correlation is computed between the SOS scores and the differences in RAD between the percentage of (women and men), (LGTBQ and straight), and (Non-white ethnicities and White ethnicities).

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.

To measure the SOS_{women} using the state-of-the-art metrics, target list W1 contained the NOI words that describe women in Table 1, 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, as described in section 3.1, and attribute list 2 a list of positive words provided by the WEFE framework. To measure the $SOS_{ethnicity}$, 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 SOS_{lgbtq}, we used the same attribute lists and target list L1, which contained NOI words that describe LGBTQ, and target list L2 which contained NOI words that describe straight and cisgender people. To measure SOS_{women}, SOS_{lgbtq}, and $SOS_{ethnicity}$ with our proposed metric, we computed the mean SOS scores of the NOI words that describe Women, LGBTQ, and Non-white ethnicities. The percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online were then acquired from the following surveys: the Rad Campaign Online Harassment Survey 2014 (Rad Campaign, 2014) where 1,000 adult Americans (aged 18+) were surveyed about being harassed online; the COX Teen Internet Safety Survey (Cox Communications Inc., 2014), where a total of 1,301 teens aged 13-17 were surveyed about being bullied online, with both surveys selected because they provide data on all the marginalised groups examined in this paper; and the online extremism and online hate survey (OEOH), collected by (Hawdon et al., 2015) from Finland (FI) (n=555), Germany (GR) (n=999), the US (n=1,033), and the UK (n=999) in 2013 and 2014, for individuals aged 15 - 30.

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. The results in Figure 2[†] show that our proposed SOS bias metric, for Chan, UD, and Glove-Twitter, has a high positive correlation with the published statistics on online abuse (RAD and COX), whereas the correlation is very small or negative for word2vec and Glove-WK. On the contrary, for the online hate and extremism surveys OEOH (US, UK, GR, and FI), our SOS bias metric for Word2Vec and Glove-WK shows a positive correlation, whereas the correlation for Glove-Twitter, UD, and Chan is negative or very small. A similar pattern is exhibited by the RNSB metric to a lesser extend. On the other hand, WEAT, RND, and ECT exhibit almost the opposite pattern, as they show a negative or very small correlation to the statistics of the surveys on online abuse (RAD and COX) for all the word embeddings, but show a high positive correlation with the statistics of the surveys of online hate and extremism OEOH (US, UK, GR, and FI).

These results suggest that our metric highlights the difference in the SOS bias between the different word embeddings, as the word embeddings that were trained on the social media datasets (Glove-Twitter, UD, and Chan) encode the online abuse towards marginalised people, while word embeddings that were trained on Google news and Wikipedia articles encode the hate and extremism against the marginalised groups shared in those sources. On the contrary, the other metrics fail to

[*]Contrary to all other 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.

[†]The correlation results for OEOH-US are similar to OEOH-UK, OEOH-GR, and OEOH-FI, so the later were omitted from the figure.

capture that difference between the word embeddings. Consequently, the results suggest that our SOS bias metric is the most reflective of the SOS bias in the different word embeddings. Additional validation of our SOS bias metric on a collection of Reddit posts is provided in Appendix A.1. The results support our findings that our SOS bias metric is reflective of the online abuse and hate experienced by marginalised groups online.

4 SOS bias and downstream tasks

In this section, we answer RQ4 through a series of experiments on the downstream task of hate speech detection. We also examined the task of offensive words categorisation in Appendix A.2.

4.1 Hate speech detection

We investigated the influence of SOS bias in the word embeddings on the task of hate speech detection by training deep learning models with an embedding layer for the detection of different types 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 contain different types of hate speech (Table 4): (i) *Twitter-racism*, a collection of Twitter messages containing tweets that are labeled as racist or not (Waseem and Hovy, 2016b); (ii) *Twitter-sexism*, Twitter messages containing tweets labeled as sexist or not (Waseem and Hovy, 2016b); (iii) Twitter-hate, containing tweets that are labeled as offensive, hateful (sexist, homophobic, and racist), or neither (Davidson et al., 2017). 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 speech 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, i.e. Women, LGBTQ, and Non-white ethnicities, thus they are representative of the examined problem.

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

Note: Positive samples refer to offensive comments

Table 4: Hate speech datasets' details.

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" were not removed, as suggested in (Elsafoury et al., 2021). All letters were lowercased, and common contractions were converted to their full forms. Finally, each dataset was randomly split into training (70%) and test (30%) sets, 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 layers 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 pre-trained 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}. For each dataset, we used 5-fold cross-validation to train and validate the model (70% and 30% of the training dataset respectively with class ratio preserved) and then test the model on the test set. We trained the models 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).

4.2 Experimental Results

Given the results for the SOS bias in the different embeddings (Table 3), we hypothesise that the deep learning models that are trained with Word2Vec 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, UD, and Chan will achieve the highest F1 scores on datasets that contain insults towards women, which are Twitter-racism and HateEval. Given that the Twitter-hate dataset contains a mixture of sexist, homophobic, and racist comments, we hypothesise that the models trained with Glove-Twitter, UD, and Chan will perform the best.The classification performance of the deep learning models with the different embedding models is reported in Table 5. The results show that for all datasets, BiLSTM outperforms MLP in terms of F1 score. In addition, results show that for the MLP model, our hypotheses hold for all four datasets, as Chan is the best performing for a dataset that contains insults towards women (HateEval), Word2Vec is the best performing on a dataset that contains insults towards other ethnicities (Twitter-racism), Glove-Twitter is the best performing on a dataset that contain insults towards women (Twitter-sexism), and UD is the best performing on Twitter-hate which contain insults towards women and the LGBTQ community. For the BiLSTM model, our hypotheses hold for three datasets, i.e., HateEval, Twitter-sexism, and Twitter-hate, as Glove-Twitter is the best performing on datasets that contain insults towards women and LGTBQ, which are found in the HateEval, Twitter-sexism, and Twitter-hate datasets. As for the Twitter-racism dataset, we hypothesised that Word2Vec would be the best performing, but instead, Glove-WK is the best performing when the BiLSTM model is used.

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

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

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

Dataset	Model	Spearman's correlation				
		WEAT	RNSB	RND	ECT	Our_metric
HateEval	MLP	**0.900**	-0.300	0.400	-0.100	0.500
	BiLSTM	0.102	-0.974	-0.461	-0.205	**0.974**
Twitter-sexism	MLP	-0.359	-0.564	-0.359	-0.615	**0.461**
	BiLSTM	-0.205	-0.102	0.153	-0.872	**0.205**
Twitter-racism	MLP	-0.900	-0.200	-0.600	-0.100	**0.100**
	BiLSTM	-0.500	**0.500**	0.200	-0.300	-0.300
Twitter-hate	MLP	**0.300**	-0.100	0	0	-0.200
	BiLSTM	**0.900**	-0.300	0.500	-0.500	0.400

Table 6: Spearman's rank 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.

To quantify our analysis of the influence of the SOS bias on the task of hate speech detection, we used Spearman's rank correlation coefficient to compute the correlation between the ranking of the mean SOS bias scores (our_metric) and the SOS bias scores as measured by WEAT, RNSB, RND, and ECT, and the ranking of F1 scores for the MLP and BiLSTM models for the different word embeddings in each examined dataset. To measure the SOS bias in the word embeddings, we used target list M1 contained the NOI words that describe the marginalised groups in Table 1 and target list N1 contained the NOI words that describe the non-marginalised groups. We used the same list of swear words described in Section 3.1 as attribute list 1 and a list of positive words, available at WEFE, as attribute list 2. We then measured the bias using the different metrics and ranked the scores in ascending order, except for ECT which is ranked in descending order because ECT scores have an inverse relationship with the level of bias.

Results in Table 6 show that our metric exhibits positive correlation with the F1 scores of the Bi-LSTM and MLP models on the HateEval and Twitter-sexism datasets. For Twitter-racism, RNSB shows the highest positive correlation with the F1-score of the Bi-LSTM model, while for the Twitter-hate dataset, WEAT shows the highest positive correlation with the F1-scores of the MLP and Bi-LSTM models. These results suggest that our SOS bias metric correlates consistently positively with the F1 scores of the deep learning models on the different datasets compared to the other metrics. Our findings in this section and in Appendix A.2 suggest that there is an influence of the SOS bias in the word embeddings on downstream tasks. It is less evident for the task of offenses categorisation but clearer for the task of hate speech detection. However, the results are not conclusive and more experiments are required.

5 SOS bias vs stereotypical social bias

To answer RQ5, we compared SOS bias, measured by our proposed metric, to stereotypical social bias, measured by state-of-the-art metrics from the literature (WEAT, RND, RNSB, and ECT), for the task of hate speech detection. We built on our findings from the previous section, assuming that the bias in word embeddings has, to some extent, an influence on the performance of the deep learning models. In this section, the comparison was performed on the task of sexism detection, thus the metrics were used to measure gender bias. The same experiment was also conducted for racial bias in Appendix A.3. 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 contains female-related words (e.g., she, woman, and mother), and Target list 2, which contains male-related words (e.g., he, father, and son), as well as two attribute lists: Attribute list 1, which contains words related to family, arts, appearance, sensitivity, stereotypical female roles, and negative words, and Attribute list 2, which contains words related to career, science, math, intelligence, stereotypical male roles, and positive words, and (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, as computed in Section 3.2 (Figure 1). For each bias metric, we ranked the bias scores for each word embedding in ascending order, except for the ECT metric that was ranked in descending order, as ECT scores have an inverse relationship with the level of bias. We then computed the Spearman's rank correlation coefficient between the gender bias of the different word embeddings, as measured by WEAT, RND, RNSB, ECT, SOS_{women}), and the F1 scores achieved by the two deep learning models on the Twitter-sexism, HateEval, and Twitter-hate datasets, using the different word embeddings (as computed in Section 4.2/Table 5). The computed Spearman's correlations are shown in Table 7.

Our results show that for HateEval and Twitter-hate, SOS_{women} has a higher positive correlation to the F1 scores of the deep learning models than the rest of the bias metrics, indicating that the SOS bias score of the different word embeddings correlates positively with the performance of the deep learning models using the word embeddings for the task of hate speech detection on these two datasets. However, for Twitter-sexism, SOS_{women} shows almost no correlation with the F1 scores of either MLP or BiLSTM. We speculate that the reason is that 66% of the Twitter-sexism dataset contains sexist tweets that are not profane, in comparison to only 40% in HateEval and Twitter-hate datasets. Our analysis showed that the gender bias scores of WEAT, ECT, RND, and RNSB metrics for the different word embeddings do not always correlate with the deep learning models' performances using the same word embeddings on the gender-relevant datasets and differ drastically from one dataset to another. The proposed SOS bias score for the different word embeddings shows a more consistent positive correlation with the F1 scores of the deep learning models using these word embeddings when profanity is used against the bias-target group. Similar results were found for racial bias, as presented in Appendix A.3. This indicates that our proposed SOS bias metric is more indicative of the sexist and racist word embeddings than the stereotypical social bias for hate speech detection.

Dataset	Model	Spearman's correlation				
		WEAT	RNSB	RND	ECT	SOS
HateEval	MLP	-0.600	0.300	0.300	-0.100	**0.800**
	BiLSTM	-0.410	-0.718	-0.307	-0.205	**0.359**
Twitter-sexism	MLP	**0.153**	-0.102	-0.205	-0.615	0.051
	BiLSTM	**0.564**	0.461	0.359	-0.872	0.05
Twitter-hate	MLP	-0.700	0.100	-0.400	0	**0.500**
	BiLSTM	-0.600	0.300	0.300	-0.500	**1**

Table 7: Spearman's rank correlation coefficient of the gender bias scores of the different word embeddings and the F1 scores of the used models for each bias metric and dataset.

6 Conclusion

In this work, we introduced the SOS bias and proposed methods to measure it, validate it, investigate its influence on downstream tasks, and compare it to stereotypical social bias. Our results show that the examined word embeddings are SOS biased and that for some of them, it has a strong positive correlation with published statistics on online abuse and extremism. However, more datasets need to be collected to provide stronger evidence, especially data from social sciences on the offenses that marginalised groups receive on social media. Our findings show that proving the influence of bias in word embeddings on the downstream tasks is not an easy task and that even though our results suggest that there is a relationship between the SOS bias and the downstream task of hate speech detection, the results are not conclusive, as there might be other factors that contributed to the performance of the examined deep learning models. Finally, our findings suggest that our proposed SOS bias metric is more indicative of the biased word embeddings in comparison to social bias for the tasks of sexism and racism detection. As future work, more experiments are required using counterfactual datasets and feature importance scores of NOI words to ensure that we understand the impact of the SOS bias in the word embeddings on the downstream tasks. Furthermore, studying the influence of particular selections of NOI words on our proposed metric will also be the focus of future work.

References

Oshin Agarwal, Funda Durupınar, Norman I. Badler, and Ani Nenkova. 2019. Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 205–211, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. 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.

Cox Communications Inc. 2014. 2014 teen internet safety survey. [Online] Accessed 13/9/2021.

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.

Sunipa Dev and Jeff M. Phillips. 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.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011*, volume WS-11-02 of *AAAI Workshops*. AAAI.

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, 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).

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. *Nordicom-Information*, 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.

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.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.

Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. 2020. "and we will fight for our race!" A measurement study of genetic testing conversations on reddit and 4chan. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 452–463. AAAI Press.

Dong Nguyen, Barbara McGillivray, and Taha Yasseri. 2017. Emo, love, and god: Making sense of urban dictionary, a crowd-sourced online dictionary. *CoRR*, abs/1712.08647.

NLTK. 2021. Nltk collocations. `https://www.nltk.org/howto/collocations.html`. Accessed: 2021-09-14.

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. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

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

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

Natalia Pietraszewska. 2013. A qualitative and quantitative analysis of selected ethnic and racial terminology present in assorted public english corpora. *Styles of Communication*, 5(1).

Rad Campaign. 2014. The rise of online harassment. [Online] Accessed 13/9/2021.

Reddit. 2021. Pushshift-reddit. `https://files.pushshift.io/reddit/comments/`. Accessed: 2021-09-14.

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

Dario Stojanovski, Gjorgji Strezoski, Gjorgji Madjarov, and Ivica Dimitrovski. 2015. Twitter sentiment analysis using deep convolutional neural network. In *Hybrid Artificial Intelligent Systems*, pages 726–737, Cham. Springer International Publishing.

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.

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.

Pierre Voué, Tom De Smedt, and Guy De Pauw. 2020. 4chan & 8chan embeddings. *CoRR*, abs/2005.06946.

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.

Steven R. Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang NLP applications. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4764–4773. European Language Resources Association.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew B. A. McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *ACM CHIL '20: ACM Conference on Health, Inference, and Learning, Toronto, Ontario, Canada, April 2-4, 2020 [delayed]*, pages 110–120. ACM.

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.

A Appendix

A.1 SOS bias validation

We compared our SOS scores to the collocations between the NOI words of marginalised groups and swear words following the work of (Pietraszewska, 2013). To generate these collocations, we used a corpus of randomly sampled 100,000 Pushshift's public Reddit collection (Reddit, 2021) comments (4 million tokens) that were posted between 2005 and 2012. Then, we used NLTK (NLTK, 2021) to find the words that co-occur the most with the NOI words and filtered them to find the co-occurrences between the NOI words w_i and the swear words o_j. The association between the acquired word pairs was measured using the pointwise mutual information (PMI). Then we computed the mean PMI for all the co-occurrences of offensive words and each of the NOI words (Equation 2). Finally, we computed the Spearman's rank correlation coefficient between the ranked mean PMI, $\overline{PMI_i}$, and the ranked SOS score $SOS_{i,we}$, for each NOI word w_i and word embedding we.

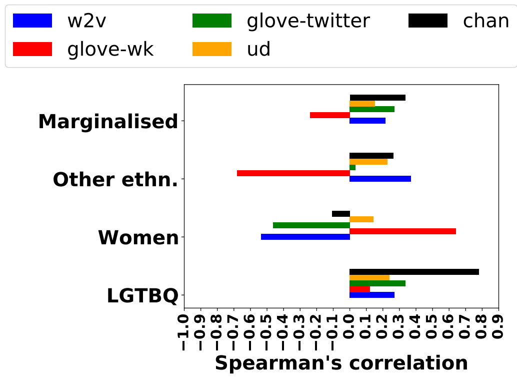

Figure 3: The Spearman's rank correlation coefficient between the ranking of SOS measure and the ranking of the mean collocation PMI.

$$\overline{PMI_i} = \frac{1}{m}\sum_{j=1}^{m} PMI(w_i, o_j) \quad (2)$$

Results in Figure 3, show a positive correlation for all the marginalised groups and most of the word embeddings, except for Glove-WK for "Other ethnicities" and Word2Vec, UD, and Chan for "Women", where a negative correlation is detected. After inspecting the "Women"-related words, where the correlation is negative, we found that they collocated with slurs that are not widely used and were not included in the used swear words list[‡]. All the NOI words in the marginalised group shows a positive correlation with all the word embeddings except for Glove-WK. We speculate that this is the case because, as shown in Figure 1 and Table 3, Glove-WK is the least biased towards "Other ethnicities".

A.2 Offensive words categorisation

We investigated the influence that the SOS bias in the word embeddings has over the downstream task of offenses categorisation. We used the Hurtlex lexicon (Zhang et al., 2020), which is a multilingual lexicon containing 8,228 offensive words and expressions, organised into 17 groups. We used words from the English lexicon that belong to the 11 groups that are related to the marginalised groups

[‡]We have not added these slurs to the swear words' list as more validation work would be required to confirm that they unambiguously belong in the list, thus risking biasing our results based on our observations.

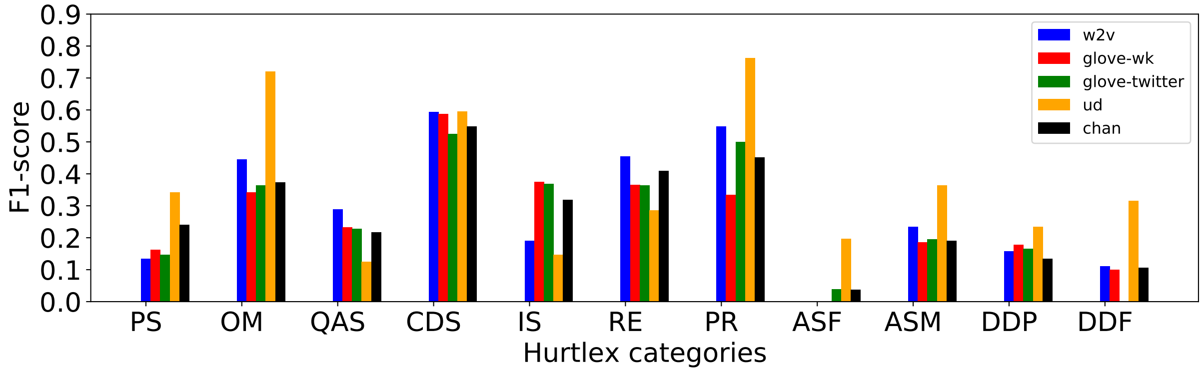

Figure 4: F1 scores for each class of the kNN model using each word embedding on the Hurtlext test set

studied in this work. The used categories are ethnic slurs (PS); words related to social and economic disadvantage (IS), descriptive words with potential negative connotations (QAS), derogatory words (CDS), felonies and words related to crime and immoral behavior (RE), male genitalia (ASM), female genitalia (ASF), words related to prostitution (PR), words related to homosexuality (OM), cognitive disabilities and diversity (DDP), and physical disabilities and diversity (DDF).

To investigate the influence that the SOS bias has on the ability of each word embedding to group together the words that belong to the same Hurtlex category, we trained a KNN model. We first removed the words in the lexicon that belong to more than one category, resulting in 5,963 offensive words in total. We then split the Hurtlex lexicon into a training (70%) and a test (30%) set, preserving the class ratio. The F1-scores achieved by the KNN model for each of the 11 classes for the test set are shown in Figure 4. A Friedman test ($\alpha = 0.05$) between the F1 scores of each data item in the test set showed that the F1 scores achieved using the examined word embeddings are significantly different. To further investigate the difference between pairs of top-scoring word embeddings, we used a Wilcoxon test ($\alpha = 0.05$). Results showed that, across all classes, UD scores significantly higher than Chan and Glove-WK, but not significantly higher than Word2Vec or Glove-Twitter. Similarly, we found that Word2Vec achieves a significantly higher F1 score than Chan and Glove-WK, but not significantly higher than Glove-Twitter. The results suggest that the UD embeddings, along with Word2Vec and Glove-Twitter, place offensive words semantically close to other words from the same Hurtlex categories, indicating that these embeddings better reflect the categorisation of terms outlined in Hurtlex. Additionally, we hypothesised that (a) Word2Vec will perform the best at classifying offensive words that are related to minorities, which are in the PS, IS, RE, QAS, and CDS classes, (b) Glove-WK will perform the best for words related to homosexuality, which are in the OM, and CDS classes, and (c) Glove-Twitter, UD, and Chan will perform best for words related to women, which are in ASF, OM, PR, and CDS classes. The results showed that our hypothesis holds for UD regarding OM, ASF, and PR and for Word2Vec regarding RE and QAS. However, for the rest of the word embeddings, our hypotheses do not hold, as Glove-Twitter and Glove-WK perform the best at classifying the words in the IS category, where Word2Vec was expected to perform the best, while Chan did not outperform any other word embeddings. Consequently, the acquired results do not provide conclusive answers to how the SOS bias in word embeddings influences the downstream task of offensive words categorisation.

A.3 Racial bias

To measure the racial bias using the state-of-the-art metrics, we used two target groups: Target group 1, which contains white people's names, and Target group 2, which contains African, Hispanic, and Asian names, and two attribute lists: Attribute list 1, which contains white people occupation names; and Attribute list 2, which contains 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 to the "Other ethnicities" category, as computed in Section 3.2 (Figure 1).

Finally, we ranked the bias scores as described in Section 5 and computed the Spearman's rank correlation coefficient between the racial bias scores of the different word embeddings and the F1 scores achieved by the two deep learning models on the Twitter-racism and HateEval datasets using the different word embeddings.

The results in Table 8 show that for Twitter-racism, SOS has the highest positive correlation with the F1 scores of the MLP model compared to the rest of the bias metrics, whereas WEAT has the highest correlation with the F1 scores of the BiLSTM model. For HateEval, SOS has the highest positive correlation with the F1-scores of the BiLSTM model compared to the rest of the bias metrics, whereas RNSB has the highest correlation with the F1 scores of the MLP model, with SOS only having a higher correlation than WEAT.

Dataset	**Model**	**Spearman's correlation**				
		WEAT	**RNSB**	**RND**	**ECT**	**SOS**
Twitter-racism	MLP	0.200	-0.900	-0.700	-0.200	**0.300**
	BiLSTM	**0.600**	-0.700	-0.100	-0.200	-0.100
HateEval	MLP	-0.200	**0.900**	0.300	0.200	0.300
	BiLSTM	-0.205	0.153	-0.718	0.205	**0.872**

Table 8: Spearman's rank correlation coefficient of the racial bias scores of the different word embeddings and the F1 scores of the deep learning models for each bias metric and dataset.