A Comparative Study on Word Embeddings in Social NLP Tasks

Fatma Elsafoury[1], Steven R. Wilson[2] 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

Abstract

In recent years, grey social media platforms, those with a loose moderation policy on cyberbullying, have been attracting more users. Recently, data collected from these types of platforms have been used to pre-train word embeddings (social-media-based), yet these word embeddings have not been investigated for social NLP related tasks. In this paper, we carried out a comparative study between social-media-based and non-social-media-based word embeddings on two social NLP tasks: Detecting cyberbullying and Measuring social bias. Our results show that using social-media-based word embeddings as input features, rather than non-social-media-based embeddings, leads to better cyberbullying detection performance. We also show that some word embeddings are more useful than others for categorizing offensive words. However, we do not find strong evidence that certain word embeddings will necessarily work best when identifying certain categories of cyberbullying within our datasets. Finally, We show even though most of the state-of-the-art bias metrics ranked social-media-based word embeddings as the most socially biased, these results remain inconclusive and further research is required.

Content Warning: As part of our experiments, we show some offensive words.

1 Introduction

Distributional word representations have been successfully used for many NLP tasks. Some of these word embeddings were pre-trained on news articles like Word2vec (Mikolov et al., 2021) or Wikipedia articles like GloVe (Pennington et al., 2021b). We use the term "informational-based" to describe these word embeddings. In recent years, there have been new word embedding models pre-trained on more informal text corpora like Twitter, 4&8 Chan and Urban Dictionary. We use the term "social-media-based" to describe those word embeddings. These informal sources contain linguistic diversity, racial slurs and forms of profanity that do not exist in formal text (Türker et al., 2016). However, these social-media-based word embeddings have not been investigated for social NLP related tasks like cyberbullying detection and social bias analysis. Our intuition that social-media-based word embeddings could be better at detecting cyberbullying comes from the examples shown in Table 1, where we display the most similar five words found by each word embeddings to the word "queer". The informational-based word embeddings return non-offensive words while social-media-based word embeddings return offensive* words. Previous research has established that word embeddings, in general, contain social biases (Garg et al., 2018; Manzini et al., 2019; Sweeney and Najafian, 2019; Bolukbasi et al., 2016; Chaloner and Maldonado, 2019). Studying social bias in word embeddings includes measuring the statistical association between certain characteristics and certain groups of people. This includes racial bias (Garg et al., 2018; Manzini et al., 2019; Sweeney and Najafian, 2019) and gender bias (Garg et al., 2018; Bolukbasi et al., 2016; Chaloner and Maldonado, 2019). Prior work has focused mainly on Word2vec, Glove-WK, and glove-twitter (Badilla et al., 2020). However, this bias has not been explored in word embed-

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

Table 1: Top 5 similar words retrieved by each of the word embeddings.

*Throughout this paper, we differentiate between the terms "offensive" and "profane": we use the term "offensive" to describe an expression that is offensive to a group of people but not necessarily profane e.g. "women belong to the kitchen" while we use the term "profane" to describe expressions like "b*tch".

dings that were pre-trained on Urban Dictionary and 4&8 Chan platforms. Since those platforms are rife with offensiveness against women and racially insensitive comments (Nguyen et al., 2017; Voué et al., 2020), this motivates our investigation into the bias in social-media-based word embeddings, especially Urban Dictionary and Chan, in comparison to informational-based word embeddings.

In this paper, we compared static word embeddings based on the datasets they were pre-trained on and not models that were used to pre-train them e.g. skip-gram. While using one model to pre-train all word embeddings on different pre-training datasets would directly show the impact of the source datasets for a particular word embedding training method, we focus our work on analyzing existing, publicly released word embeddings which are often used in other downstream tasks in order to better understand the impact of using these embeddings. We examined static word embeddings instead of contextual word embeddings as they are still widely used in NLP tasks and there have not been any released contextual word embeddings pre-trained on datasets like Urban Dictionary or Chan, and pre-training these models from scratch is computationally expensive.

We set out to answer the following research questions: 1) What is the performance of the different word embeddings on offences categorisation?. 2) What is the performance of the different word embeddings on the task of cyberbullying detection? Can we use certain word embeddings to detect certain offensive categories within cyberbullying-related datasets? 3) Are social-media-based word embeddings more socially biased than informational-based word embeddings? To answer the first research question, we used the different word embeddings to categorize terms from a popular lexicon of English offensive language. Then we compared the performance of the social-media-based word embeddings and the informational-based word embeddings using statistical significance tests. Answering our first research question should help in finding out whether social-media-based word embeddings are significantly better than informational-based word embeddings at learning the semantic relationship between terms that belong to the same group of offences. We answer our second research question through a series of experiments where we used each word embedding to automatically detect cyberbullying in cyberbullying-related datasets and to detect different types of cyberbullying within each dataset. We used a statistical significance test to compare the performance of the social-media-based word embeddings and the informational-based word embeddings. Answering the second research question will help us to find out if social-media-based word embeddings improve the performance on the task of cyberbullying detection in comparison to informational-based word embeddings and to find out the ability of certain pre-trained word embeddings to detect certain types of cyberbullying. Finally, to answer our last research question and to find out which word embeddings are more socially biased, we used the state-of-the-art metrics from the literature to measure gender and racial bias in each word embedding and compared the bias scores in the social-media-based word embeddings and the informational-based word embeddings.

The contributions of this paper are: **(a)** We demonstrate that social-media-based word embeddings are better at categorizing offensive words and that social-media-based word embeddings outperform informational-based word embeddings on cyberbullying detection. **(b)** Our findings show no evidence that certain word embeddings are better than others at detecting certain offensive categories within the examined cyberbullying-related datasets. **(c)** Our results show no strong evidence that social-media-based word embeddings are more socially biased than informational-based word embeddings. We share our code with the community to reproduce our results and allow more investigation [†].

2 Related work

Recent word embeddings pre-trained on data from social media platforms have been released in the community. For example, Urban Dictionary word embeddings that was pre-trained on words and definitions from the Urban Dictionary website (Wilson et al., 2020) using the FastText framework, Chan word embeddings that was pre-trained on 4&8 Chan websites using Continuous Bag-of-Words algorithm (CBOW) (Voué et al., 2020), and a version of Glove pre-trained on Twitter data (Pennington et al., 2021a). Even though there is evidence from the literature that the data that was used in pre-training these word embeddings contain offensive-

[†] https://github.com/efatmae/Comparative_analysis_word_embeddings_on_social_NLP_tasks

ness and racially insensitive comments (Nguyen et al., 2017; Papasavva et al., 2020), they have not been investigated for social NLP tasks. For example, investigating the impact of social-media-based word embeddings on the task of cyberbullying detection or analysing the social bias in the social-media-based word embeddings.

Using social-media-based word embeddings could improve cyberbullying detection as they may be able to identify some offensive words or forms of profanity that are not captured by informational-based word embeddings. Comparative studies on word embeddings and deep learning models have been done for biomedical natural language processing (Wang et al., 2018) and for text classification, (Wang et al., 2020), but there have been very few similar comparative studies for the task of cyberbullying detection. Jain et al. (2021) reviewed the literature on different word embeddings: CBOW, Skip-gram, ELMo, GloVe and fastText, and then tested them with a neural networks model on hate speech detection task. They show that ELMo is the best performing followed by fastText and GloVe. However, they do not include social-media-based word embeddings like Urban Dictionary or Chan. Elsafoury et al. (2021) have shown that word embeddings pre-trained on Urban Dictionary, and Twitter outperforms embeddings like Word2vec and Glove-Wikipedia on the task of cyberbullying detection. However, they do not compare the ability of the different word embeddings to categorize offensive words or to detect different categories of offences within cyberbullying datasets.

Additionally, The research has shown that word embeddings are biased. Among the most common methods for quantifying bias in word embeddings are the word embedding association test (WEAT), the relative norm distance (RND), The relative negative sentiment bias (RNSB), and The embedding coherence test (ECT). For the WEAT metric, 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. As for the RND metric, 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). As for the RNSB metric, the authors trained a logistic regression model on the word vectors of unbiased labelled 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 the ECT metric, 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 bais metrics have been used to measure the bias in Word2vec(Caliskan et al., 2017; Garg et al., 2018; Sweeney and Najafian, 2019; Dev and Phillips, 2019), Glove-WK (Dev and Phillips, 2019; Sweeney and Najafian, 2019), Glove-Twitter (Dev and Phillips, 2019). Even though research has shown that the upstream data used to pre-train the social-media-based word embeddings, especially Urban Dictionary and Chan, are full of racial slurs and profanity (Nguyen et al., 2017; Voué et al., 2020), none of these studies measured the social bias in Urban Dictionary or Chan word embeddings. In this paper, we run a series of experiments to fill the mentioned gaps in the literature and to answer our research questions.

Word embedding	Pre-training data	Type
Word2Vec	Google news articles	informational-based
Glove-Wikipedia	Wikipedia articles	informational-based
Glove-Twitter	Twitter messages	social-media-based
Chan	Text from 4&8 Chan	social-media-based
Urban Dictionary	Text from Urban Dictionary	social-media-based

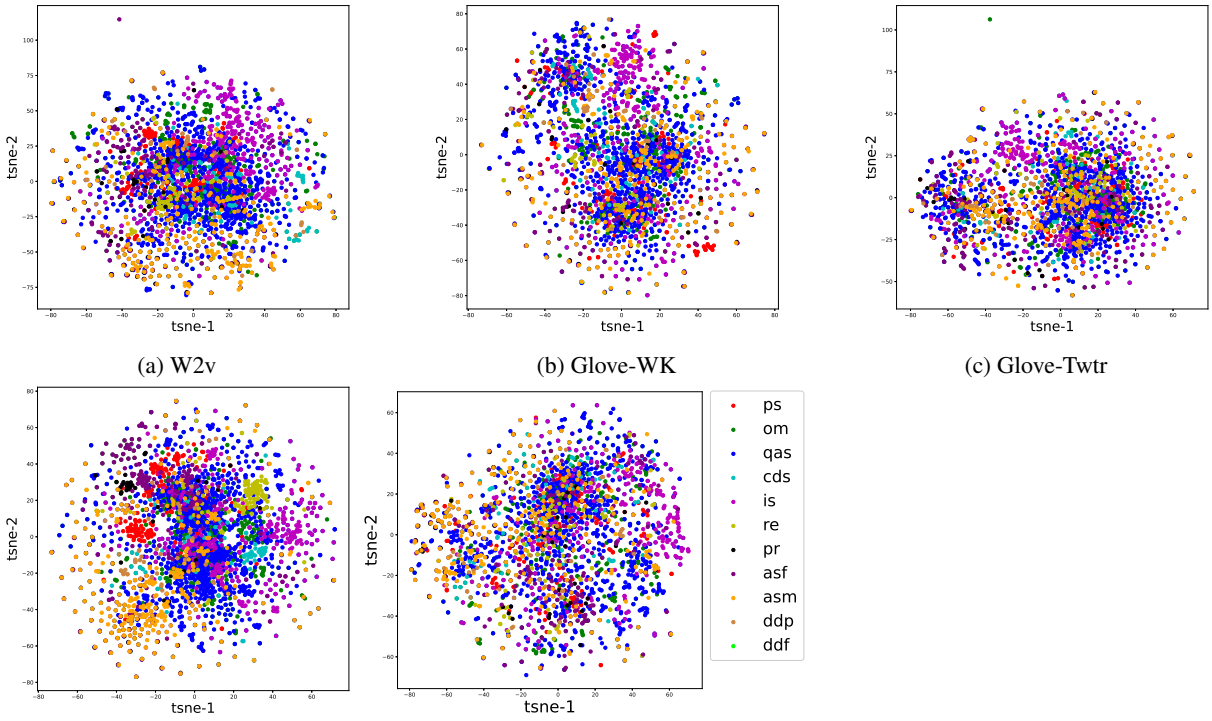
Table 2: Word embedding models used in the paper.

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

Table 3: Hurtlex categories used in this paper.

3 Offenses categorization

In this paper, we used the word embedding models that are summarized in Table 2. To answer our research questions, we used the English offensive categories introduced in Hurtlex lexicon (Zhang et al., 2020), which is a multilingual lexicon containing



(a) W2v (b) Glove-WK (c) Glove-Twtr

(d) UD (e) Chan

Figure 1: t-SNE of the different word embeddings of the words that belong to different groups in Hurtlex lexicon.

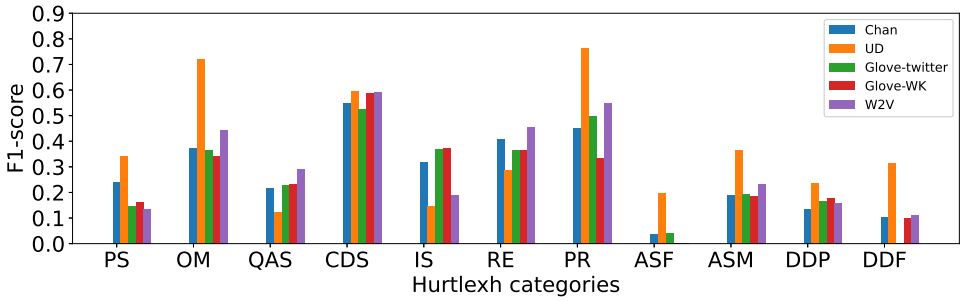


Figure 2: F1 scores of the KNN model with the different word embeddings on Hurtlext test set.

8228 offensive words and expressions, which are organized into 17 groups. We only used words that belong to 11 groups because they are related to the types of cyberbullying found in our datasets. The used categories are summarized in Table 3. We extracted the word vectors, using the different word embeddings described in Table 2, for each word in those 11 groups and projected them into a two-dimensional space using t-SNE (van der Maaten and Hinton, 2008) as shown in Figure 1. The plot shows words from some Hurtlex categories clustered better in some cases, especially, PS, PR, and ASM with Urban Dictionary. To quantitatively investigate the ability of the different word embeddings to group the words that belong to the same Hurtlex category, we used a KNN model. We first removed the words in the lexicon that belong to more than one category, which resulted in 5963 offensive words. We then split Hurtlex lexicon into training (70%) and test (30%) sets with class ratio preserved. Next, in order to understand if the neighbors of a given word typically belong to the same class as that word, we used the trained KNN model to predict the category of each word embedding in the test set based on proximity to embeddings from the training set. We measured the F1-scores and plot them in Figure 2. To answer our first question, our results show that for most of Hurtlex categories, PS, OM, PR, ASF, ASM, DDP and DDF, Urban Dictionary is the best performing, meaning that it was the best at grouping together the words that belong to these categories. For QAS

and RE, Word2vec is the best performing and for IS, Glove-Wikipedia and Glove-twitter are the best performings. For CDS, all the word embeddings are performing similarly with Urban Dictionary embedding being the best performing by a small margin. We speculate that these results stem from the fact that the Urban Dictionary is pre-trained on words and definitions that are of insulting nature in general, and to women and minorities specifically, so it is better at finding more profanity related to these categories: PS, OM, PR, ASF, ASM, DDP and DDF. Word2vec, on the other hand, is better at clustering the word vectors that are related to felonies and words related to crime and immoral behaviour (RE) and words with potential negative connotations (QAS). That may be due to its pre-training on news articles, which sometimes report on crimes. Using a Friedman significance statistical test (Zimmerman and Zumbo, 1993) ($\alpha = 0.05$) between the F1 scores of each data item in the test set, we found that the F1 scores achieved by the word embeddings are significantly different. To further investigate the difference between pairs of top-scoring word embeddings, we use a Wilcoxon test (Zimmerman and Zumbo, 1993) ($\alpha = 0.05$). We found that, across all categories, Urban Dictionary scores significantly higher than Chan and Glove-Wikipedia but not significantly higher than Word2vec or Glove-Twitter. Similarly, we found that Word2vec achieves a significantly higher F1 score than Chan and Glove-Wikipedia, but not significantly higher than Glove-Twitter. The results suggest that the Urban Dictionary 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 categorization of terms outlined in Hurtlex.

4 Cyberbullying detection

In the light of our earlier results presented in Figure 2, we make two hypotheses: (1) social-media-based word embeddings will perform better than informational-based embeddings on the task of cyberbullying detection. (2) Certain word embeddings will perform better at detecting certain offensive categories within our cyberbullying-related datasets. Specifically, we expect that Urban Dictionary embeddings might perform the best on the examples in the datasets containing PS, OM, PR, ASF, ASM, DDP and DDF categories; Word2vec embeddings to perform the best on examples containing RE and QAS; and for the CDS category, we expect all the models to perform similarly. To test our hypotheses and answer our second research question, we compared the performance of the different word embeddings when used to initialize the embedding layer of a deep learning model trained on the following datasets.

4.1 Cyberbullying datasets

We used five cyberbullying-related datasets from several social media sources that contain different types of cyberbullying: (i) *Twitter-Racism*, a collection of Twitter messages containing tweets that are labelled as racist or not (Waseem and Hovy, 2016); (ii) *Twitter-Sexism*, Twitter messages containing tweets labelled as sexist or not (Waseem and Hovy, 2016); (iii) *HateEval*, a collection of tweets containing hate speech against immigrants and women in Spanish and English (Basile et al., 2019). We used only the English tweets; (iv)*Kaggle* (Kaggle, 2012), a dataset that contains social media comments that are labelled as insulting or not; and (v) *Jigsaw*, a collection of Wikipedia Talk Pages comments which have been labelled by human raters for toxicity (Jigsaw, 2018). The datasets 'statistics are described in Table 4.

To pre-process the datasets, we removed URLs, user mentions, and non-ASCII characters; All letters were lowercased; common contractions were converted to their full forms. We also removed English stop words, as proposed in (Agrawal and Awekar, 2018). However, second-person pronouns like "you", "yours" and "your", and third-person pronouns like "he/she/they", "his/her/their" and "him/her/them" were not removed because we noticed in our datasets that sometimes, profane words on their own, e.g. "f**k", are not necessarily used in an offensive way, while their combination with a pronoun, e.g. "f**k you", is used to insult someone. For Twitter datasets, we also removed the retweet abbreviation "RT". Each dataset was randomly split into training (70%) and test (30%) sets with preserved class ratios. Additionally, to find out the different categories of offences within each cyberbullying dataset, we filtered the datasets using the words in the Hurtlex lexicon. Then we sorted the data items in each dataset into the 11 Hurtlex categories based on the words present in the data items. Those that contain a mix of words from multiple Hurtlex categories were grouped in a Mixed

category, and all the data items that do not contain any Hurtlex words were placed in a No-Hurtlex category. The results show that for all the datasets, the majority of data items contain words that do not belong to any Hurtlex category (No-hurtlex) with a percentage range from 40% to 66%. The second most present category in all the datasets is the Mixed category where the data items contain words from multiple Hurtlex categories with percentages ranging from 5% to 25%. For the data items that contain words from only one Hurtlex category, the datasets, are less than 10% except for the CDS category where the percentage is less than 20%. When we investigated the distribution of the different categories in the Mixed group, we found a similar distribution of the 11 categories in all the datasets with the majority belonging to the CDS category. When we investigated the data items in the No-Hurtlex category, we found some non-profane form of offensiveness.

4.2 Model settings

We used a Bi-directional LSTM (Schuster and Paliwal, 1997), with the same architecture as in (Agrawal and Awekar, 2018), who used RNN models to detect cyberbullying. To this end, we first used the Keras tokenizer (Tensorflow.org, 2020) to tokenize the input texts, using a maximum input length of 64 (maximum observed sequence length in the dataset) for the HateEval and Twitter datasets and 600 for the Kaggle and Jigsaw datasets (due to computational resource limitations). A frozen embedding layer, based on a given pre-trained word embedding model, was used as the first layer and fed to the Bi-LSTM model. To avoid over-fitting, we used L2 regularization with an experimentally determined value of 10^{-7}. The model was then trained for 100 epochs with a batch size of 32, using the Adam optimiser and a learning rate of 0.01.

4.3 Results

To answer the first part of our second research question, we analysed the overall performance of each word embeddings on each dataset, the "Average" column in Table 5, individually and across all the datasets. We used Friedman statistical significance test (Zimmerman and Zumbo, 1993) ($\alpha = 0.05$) to compare the F1-scores of each word embeddings for the 13 categories (PS, OM, QAS, CDS, IS, RE, PR, ASF, ASM, DDP, DDF, No-hurtlex and Mixed) in each dataset. Our results show that social-media-based word embeddings gave the best results for four out of five datasets: HateEval, Kaggle, Twitter-racism and Jigsaw-toxicity. For the HateEval dataset, performance across all the categories is at its best when Glove-Twitter, social-media-based, was used with an average F1 score of 0.620. However, the results across all the categories are not significantly better than the rest of the word embeddings with $p - value > 0.05$. Glove-Twitter also resulted in the highest average F1 score at 0.519, across all the categories on the Jigsaw-toxicity dataset which is significantly better for all the categories with $p - value < 0.05$. The best performing word embeddings on the Kaggle dataset is also the social-media-based word embeddings, Chan, with the average F1-score of 0.727 across all the categories with the results significantly better than the rest of the word embeddings for all the categories with $p - value < 0.05$. Urban Dictionary embeddings, social-media-based, gave the best results on the Twitter-racism dataset with the average F1 score of 0.663 across all the categories. These results are significantly better with $p - value < 0.05$. The informational-based word embeddings, Glove-Wikipedia, gives a significantly better average F1-score of 0.699 across all the categories on the Twitter-sexism dataset with $p - values < 0.05$. Overall, we found that although social-media-based word embeddings outperform others on four out of five datasets, the difference is only significant in three cases.

To answer the second part of the second research question, we analysed the results across the different types of cyberbullying in the datasets, we computed the mean F1-score achieved by each word embedding for each category across all datasets. When we compared the mean F1-score achieved by each word embedding for each category across all datasets using a Friedman significance statistical test ($\alpha = 0.05$), we found no significance for any of the 13 categories (PS, OM, QAS, CDS, IS, RE, PR, ASF, ASM, DDP, DDF, No-hurtlex and Mixed). This might occur because there is

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

Table 4: Cyberbullying dataset statistics. Pos. is the percentage of positive (bullying) comments. Avg. is the average number of words per comment. Max. is the maximum number of words in a comment.

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

Table 5: Binary F1-scores of the Bi-LSTM of each word embeddings on the different types of cyberbullying within each dataset and on the average F1 score across all the types. "Average" is the average F1 score for each datasets across all the 13 categories.

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

Table 6: The bias scores of the different word embeddings are measured using different metrics (higher scores indicate stronger bias). We report the ranking of the bias score and the actual bias score between brackets. **Bold** text represents the most biased.

no clear connection between the ability of word embeddings to cluster the Hurtlex categories and their performance on texts that contain the same offensive words in cyberbullying related datasets. Alternatively, due to the very small percentages of these categories in our datasets, it is possible that we could not get a reliable enough indication of the performance of each word embedding model on each category. More analysis and experiments with larger datasets where these categories are more prevalent are needed to fully understand the results.

5 Social bias

In this section, we answer our third research question by measuring the social bias in the different word embeddings. We studied two types of social bias: gender bias and racial bias. We hypothesise that social-media-based word embeddings, especially Urban Dictionary and Chan, are more socially biased than informational-based based word embedding. We used the WEFE framework (Badilla et al., 2020) to measure the gender bias and the racial bias in the different word embeddings using the state-of-the-art bias metrics from the literature: WEAT, RNSB, RND, and ECT. To measure the gender bias, we follow the methodology proposed in the original paper (Caliskan et al., 2017) using the WEFE framework (Badilla et al., 2020). We used 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. Then, we measured

the average gender bias scores across the different attribute lists for each word embedding using the various metrics. Since the different metrics use different scales, we follow the work suggested in (Badilla et al., 2020) to rank 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. Similarly, to measure the racial bias we follow the methodology proposed in (Garg et al., 2018) using the WEFE framework. 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's occupation names; and Attribute list 2, which contains African, Hispanic, and Asian people's occupations. 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). Finally, we ranked the bias scores.

The results reported in Table 6 show variations between the different bias metrics. The WEAT bias metric does not support our hypothesis with Word2vec and Glove-WK being ranked as the highest two biased word embeddings regarding gender and racial biases. On the other hand, The RNSB, RND, and ECT metrics give us mixed results. As RNSB ranked Chan and Glove-WK as the highest two biased word embeddings regarding gender bias and Chan and Urban Dictionary as the highest two biased word embeddings regarding racial bias. While RND ranked Chan and Glove-WK as the highest two biased word embeddings regarding gender and racial bias. As for ECT, the metric ranked Chan and Word2vec as the highest biased embeddings regarding gender and racial bias. The results suggest that even though according to most of the metrics (RND, RNSB and ECT), the most biased word embeddings for racial and gender bias are Urban Dictionary and Chan, which supports our hypothesis, there is no consistent evidence that social-media-based word embeddings are more biased than informational-based-word embeddings. We speculate that this is the case because social bias takes different forms some include profanity and slurs which are the cases where social-media-based word embeddings are ranked the highest biased. While some times social bias takes non-offensive forms which are the cases when Glove-WK was ranked the second most biased word embeddings.

6 Conclusion

The work in this paper was motivated by the release of the new social-media-based word embeddings. We ran a series of experiments to compare social-media-based word embeddings and informational-based word embeddings regarding two social NLP tasks: cyberbullying detection and social bias analysis. We found that social-media-based word embeddings are better than informational-based embeddings at categorizing offensive words. This suggests that social-media-based word embeddings might be useful for expanding queries to collect future cyberbullying datasets. We also found that social-media-based word embeddings performed better at the task of cyberbullying detection than informational-based word embeddings. Our results also show that although some word embeddings are better at categorizing offensive words in the Hurtlex categories, these same embeddings do not necessarily perform better at detecting the corresponding offensive categories within our datasets. Hence, there is no evidence that certain word embeddings are better at detecting certain types of cyberbullying.

Our results also show that even though the different bias metrics don't agree on the ranking of the word embeddings regarding social bias, most of the bias metrics (RNSB, RND, and ECT) agree that Chan and Urban Dictionary are the highest ranked biased word embeddings regarding gender and racial bias. However, the second highest biased word embeddings is Glove-WK which is not social-media-based which means that social-media-based word embeddings are not necessarily more socially biased than informational-based word embeddings.

Our findings raise questions about some common methods currently used to detect cyberbullying and to measure social bias in word embeddings. As our findings show that state-of-the-art bias metrics did not agree on the rankings of the most biased word embeddings. Additionally, our findings show that profanity is an important feature that should be used in addition to other features to develop more reliable models to detect cyberbullying and to reveal the social bias in the different word embeddings. Future work should investigate the relationship between the bias in the word embedding and the performance of these word embeddings on cyberbullying detection.

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.

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.

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.

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.

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.

Fatma Elsafoury, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. 2021. When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access*.

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.

Minni Jain, Puneet Goel, Puneet Singla, and Rahul Tehlan. 2021. Comparison of various word embeddings for hate-speech detection. In *Data Analytics and Management*, pages 251–265, Singapore. Springer Singapore.

Jigsaw. 2018. Detecting toxic behaviour in wikipedia talk pages. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data. Accessed: 2021-04-07.

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

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.

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.

Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. *CoRR*, abs/2001.07487.

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.

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

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.

İlker Türker, Eftal Şehirli, and Emrullah Demiral. 2016. Uncovering the differences in linguistic network dynamics of book and social media texts. *SpringerPlus*, 5(1):1–18.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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

Congcong Wang, Paul Nulty, and David Lillis. 2020. A comparative study on word embeddings in deep learning for text classification. In *NLPIR 2020: 4th International Conference on Natural Language Processing and Information Retrieval, Seoul, Republic of Korea, December 18-20, 2020*, pages 37–46. ACM.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul R. Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *J. Biomed. Informatics*, 87:12–20.

Zeerak Waseem and Dirk Hovy. 2016. 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.

Donald W Zimmerman and Bruno D Zumbo. 1993. Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks. *The Journal of Experimental Education*, 62(1):75–86.