# Darkness can not drive out darkness: Investigating Bias in Hate Speech Detection Models

Fatma Elsafoury
School of Physics, Engineering, and Computing
The University of The West of Scotland
fatma.elsafoury@uws.ac.uk

## Abstract

It has become crucial to develop tools for automated hate speech and abuse detection. These tools would help to stop the bullies and the haters and provide a safer environment for individuals especially from marginalized groups to freely express themselves. However, recent research shows that machine learning models are biased and they might make the right decisions for the wrong reasons. In this thesis, I set out to understand the performance of hate speech and abuse detection models and the different biases that could influence them. I show that hate speech and abuse detection models are not only subject to social bias but also to other types of bias that have not been explored before. Finally, I investigate the causal effect of the social and intersectional bias on the performance and unfairness of hate speech detection models.

## 1 Introduction

Over the last decade, there have been attempts to use machine learning models (Dinakar et al., 2011; Dadvar et al., 2014; Rafiq et al., 2015; Waseem and Hovy, 2016a; Raisi and Huang, 2017; Agrawal and Awekar, 2018a; Kumar et al., 2019; Pavlopoulos et al., 2019; Mozafari et al., 2019; Yadav et al., 2020; Paul and Saha, 2020) for the task of hate speech and abuse detection. However, those studies focused mainly on enhancing models' performance, without providing any insight into the models' inner workings.

In recent years, the research community started to pay more attention to machine learning models' explainability and the biases in these models and the datasets. Wagner et al. (2021) describe the term *algorithmically infused societies* as the societies that are shaped by algorithmic and human behavior. The data collected from these societies carry the same bias in algorithms and humans, like population bias and behavioral 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 biases (Garg et al., 2018; Manzini et al., 2019; Sweeney and Najafian, 2019), gender biases (Garg et al., 2018; Bolukbasi et al., 2016; Chaloner and Maldonado, 2019), and personality stereotypes (Agarwal et al., 2019).

Recent research in social science explains that using racial slurs and third person profanity goes beyond offending individuals or groups of people and that it actually aims at stressing on inferiority of the identity of marginalized groups (Kukla, 2018). However, the research on bias in NLP have not payed attention to how this type of offensive stereotyping being encoded in machine learning models that are trained on data from social media. So I introduce systematic offensive stereotyping (SOS) bias which includes associating offensive terms to different groups of people, especially marginalized people, based on their ethnicity, gender, or sexual orientation. On the other hand, studies that focused on the same type of bias in hate speech detection models studied it within hate speech datasets (Dixon et al., 2018; Waseem and Hovy, 2016b; 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, the proposed methods to study social biases like gender bias in word embeddings focused on studying the statistical association between words that describe women e.g., wife, mother, sister, girl, woman, and words related to femininity e.g. nurturing, sensitive, and emotional (Caliskan et al., 2017; Garg et al., 2018; Sweeney and Najafian, 2019; Dev and Phillips, 2019). However, social science literature has shown that femininity differs in conceptualization among White and black people (Giddings, 2006; Rosenfield, 2012). Additionally, the claim that the bias found in the word

embeddings influence the NLP downstream tasks has not been proven (Blodgett et al., 2020). A few studies have used statistical correlation to show that influence (De-Arteaga et al., 2019). However, correlation is not causation and causal inferences have not been used to understand the influence of bias that exists in word embeddings, on the downstream task of hate speech detection.

The limitations enlisted here could have negative implications as hate speech detection models might learn to associate marginalized groups with extremism and abuse. As a result, these models that were supposed to provide a protective environment for the marginalized people to express themselves are the ones that could lead to silencing them or flagging their content as inappropriate. In this thesis, I aim to understand and investigate the performance and the biases of hate speech and abuse detection models through achieving the following research goals: 1) Understand the performance of state-of-the-art hate speech and abuse detection models. 2) Inspect other biases than social stereotypical bias in commonly used static word embeddings. 3) Investigate intersectional bias in contextual word embeddings and the causal effect of social and intersectional bias on the task of hate speech detection.

## 2 Literature review

### 2.1 Hate speech detection

In the literature on hate speech and abuse detection, there is a lack of clear distinction between hate speech and related concepts like online abuse (Elsafoury et al., 2021). There are different definitions of online abuse but most of them can be summarized as "*one form or another of insulting, spread using mobile or internet technology*" (Elsafoury et al., 2021). On the other hand, Fortuna et al. studied hate speech in the literature in relation to four dimensions: physical violence encouragement, targets, attack language, and humorous hate speech and introduced the following definition "*a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used*" (Fortuna and Nunes, 2018). I could distinguish between online abuse and online hate speech through the target of the attack. If the target is an individual then it is online abuse but if the target is a group of people then it is online hate speech. Since I'm investigating bias which is related to groups of people, so in this thesis, I focus on hate speech detection rather than abuse detection except for the first research goal where online abuse datasets are used.

Different approaches have been developed to detect hate speech and abuse detection from social media including rule-based, conventional, deep learning, and attention-based machine learning models (BERT) (Elsafoury et al., 2021). These studies have shown that BERT outperformed all the other models on the task of hate speech and abuse detection (Paul and Saha, 2020; Mozafari et al., 2019). However, none of them explain why. In the last few years, there have been published studies on the analysis of BERT's attention weights on the GLUE tasks (Kovaleva et al., 2019; Rogers et al., 2021; Sun and Lu, 2020; Vashishth et al., 2019; Serrano and Smith, 2019) but none of them were employed for the task of hate speech and abuse detection. Inspired by this research, one of my research goals in this thesis is to gain a better understanding of BERT's strong performance on the task of hate speech and abuse detection.

### 2.2 Bias in word embeddings

The term *bias* is defined and used in many different ways (Olteanu et al., 2019). Most of the studies that measure bias in NLP use 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 (static word embeddings), the most commonly used methods for quantifying bias are WEAT , RND, RNSB, and ECT (Badilla et al., 2020). 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) ex-

tracted 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 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 are unreliable as explained by (Antoniak and Mimno, 2021). Since I am interested in measuring the systematic offensive stereotypes of different marginalized groups based on race and sexual orientation, these metrics would fall short of my 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 my case, I am interested in a variety of offensive language targeted at different marginalized groups. Additionally, the literature on bias in word embeddings claims that it influences downstream tasks, like translation, classification, and text generation. Still, these claims have not yet been tested (Blodgett et al., 2020). In this thesis, I aim to address these limitations by introducing the systematic offensive stereotyping (SOS) bias, proposing a method to measure it, and investigating the statistical association between the SOS bias and the task of hate speech detection.

### 2.3 Intersectionality of bias

Intersectionality as a term is coined by Kimberle Crenshaw (Crenshaw, 1989) to describe that Black women experience a different type of bias other than the ones experienced by White women and Black men. She states that "*This intersectional experience is greater than the sum of racism and sexism, any analysis that does not take intersectionality into account can not sufficiently address the particular manner in which Black women are subordinated*" (Crenshaw, 1989). Ever since there has been increasing research on intersectionality in social sciences. For example, European American people associate femininity with characteristics like submissiveness, nurturing, sensitivity, and emotional expressiveness. On the contrary, for African American people, femininity incorporate paid work and achievement. African American people conceptualize gender as flexible with greater gender role equality and less traditional attitudes towards women's roles than European American people (Giddings, 2006; Rosenfield, 2012). Similarly, O'Brien et al., show that African American women are more likely to major in STEM fields in comparison to European American women. They also found that African Americans had a weaker implicit gender-STEM stereotype than European Americans (O'Brien et al., 2015). These Examples show that the methods used in the literature to measure the gender bias in word embeddings (WEAT, RND, and ECT) measure the gender bias that European American women suffer from "White gender bias" which does not reflect the experience of women of color especially African American women.

A few studies focus on the intersectionality of bias in pre-trained contextual word embeddings (Guo and Caliskan, 2021; Tan and Celis, 2019; Lepori, 2020). These studies have used seed words from the literature for their tests without mitigating for their limitations as specified by (Antoniak and Mimno, 2021). The limitations include the lack of motivation behind choosing and the lack of coherence among the words that describe the same group of people like using people's names to infer their ethnicity or race. Additionally, the inspected intersectional biases have not been tested for their influence on downstream tasks. For example, (Kim et al., 2020) investigated the intersectional bias in hate speech datasets again without analyzing their influence on the model's outcome.

In this thesis, I aim to mitigate this limitation by creating a new bias dataset and propose a method to measure interseactional bias in contextual word embeddings. Additionally, I am going to investigate the causal influence of the studied intersectional bias on the task of hate speech detection.

### 2.4 Causality in NLP

As mentioned earlier the research community has mainly focused on measuring bias in word embeddings without understanding how this bias influences the downstream NLP tasks. Even the few studies that investigated that influence, have re-

| Dataset | Samples | Positive samples | |
|---|---|---|---|
| Kaggle-insults | 7425 | 35% | (Kaggle, 2012) |
| Twitter-sex | 14742 | 23% | (Waseem and Hovy, 2016a) |
| Twitter-rac | 13349 | 15% | (Waseem and Hovy, 2016a) |
| HateEval | 12722 | 42% | (Basile et al., 2019) |
| Twitter-hate | 5569 | 25% | (Davidson et al., 2017) |
| WTP-agg | 114649 | 13% | (Wulczyn et al., 2017) |
| WTP-tox | 157671 | 10% | (Wulczyn et al., 2017) |

Table 1: Dataset statistics

| Dataset | LSTM | Bi-LSTM | BERT(FT) |
|---|---|---|---|
| Kaggle-insults | 0.6420 | 0.653 | 0.768 |
| Twitter-sex | 0.6569 | 0.649 | 0.760 |
| Twitter-rac | 0.6400 | 0.678 | 0.757 |
| WTP-agg | 0.7110 | 0.679 | 0.753 |
| WTP-tox | 0.7230 | 0.737 | 0.786 |

Table 2: F1-scores achieved for each dataset

lied on statistical correlations. For example, De-Arteaga et al., measure the correlation between the true positive rates gap between genders in the task of occupation classification and the existing gender imbalances in those occupations (De-Arteaga et al., 2019).

Given that correlation is not causation, there has been a recent trend in NLP that uses causal inference to understand the influence of different concepts on different NLP tasks (Feder et al., 2021a). Some of these studies have focused on understanding the causal inference of concepts (e.g. social bias in the datasets) on the task of text classification using counterfactual causal inference (Feder et al., 2021b; Qian et al., 2021; Elazar et al., 2021). Others have focused on using causal inferences to understand the influence of some concepts (e.g. syntax representation, and social biases in pre-trained word embeddings) on tasks like consistency with English grammar (Ravfogel et al., 2021; Tucker et al., 2021). However, causal inference methods have not been used to investigate the influence of bias in pre-trained word embeddings on hate speech. In this thesis, I aim to fill that research gap by using counterfactual causal inference to measure that influence and to measure how harmful that influence is on the task of hate speech detection.

## 3 Proposed Methods

In this section, I describe the proposed methods to achieve my research goals and the outcomes of the research goals that have been achieved. The datasets used in the experiments discussed in sections 3.1 and section 3.2 are described in Table 1.

### 3.1 Research objective 1

To achieve my first research goal, I started with reviewing the literature on hate speech and abuse detection models including the most used ML models, and datasets. Then, weused BERT in comparison to RNN models on the task of hate speech and abuse detection using some. or fine-tuning, BERT was trained for 10 epochs with a batch size of 32 and a learning rate of $2e^{-5}$, as suggested in (Devlin et al., 2019). The sequence length parameter changed across datasets depending on their maximum token length. For the Twitter-sexism and Twitter-racism datasets, a sequence length of 64 was used because it is the closest to the maximum observed sequence length in the dataset, while 128 was used for the rest because it is the maximum I could use due to available computational resources limitations. A single linear layer was added on top of the pooled output of BERT for sentence classification. I also used LSTM (Hochreiter and Schmidhuber, 1997) and Bi-directional LSTM (Schuster and Paliwal, 1997), with the same architecture as in (Agrawal and Awekar, 2018a), who used RNN models to detect cyberbullying. To this end, I first used the Keras tokeniser (Tensorflow.org, 2020) to convert the text into numerical vectors (each integer being the index of a token in a dictionary) with a maximum length of 600 (the maximum I could use due to computational resources limitations) for the Kaggle and WTP datasets and 41 (maximum observed sequence length in the dataset) for the Twitter datasets. A trainable embedding layer was used as the first hidden layer of the LSTM and Bi-LSTM-based networks, with an input size equal to the number of unique tokens of the dataset after pre-processing and an output size of 128. The two models were then trained for 100 epochs with a batch size of 32, using the Adam optimiser and a learning rate of 0.01 which is the default of the Keras Optimiser. The results show that BERT outperforms other commonly used deep learning models on multiple hate speech and abuse-related datasets achieving the highest F1 (Table 2).

I built on these results by analyzing the performance of BERT to understand the reason behind BERT's good performance (Elsafoury). To achieve that I first examined how fine-tuning affects BERT's attention weights, the results show that there is a difference in attention weights' patterns between BERT with and without fine-tuning. Then, to investigate the role of attention weights

| Dataset | No. tokens | PCC (attention vs importance) | PCC (attention vs no. occurrences) | PCC (importance vs no. occurrences) |
|---|---|---|---|---|
| Twitter-Sexism | 3878 | 0.108 | -0.047 | -0.002 |
| Twitter-Racism | 3991 | 0.056 | -0.015 | -0.002 |
| Kaggle-Insults | 4452 | 0.171 | -0.023 | -0.004 |
| WTP-Aggression | 4457 | 0.125 | -0.101 | -0.009 |
| WTP-Toxicity | 4524 | 0.163 | -0.076 | -0.011 |

Table 3: PCC between mean attention weights of fine-tuned BERT, mean absolute feature importance and number of occurrences per token

of a fine-tuned BERT in the model's performance, I compared the mean feature importance score of individual tokens, obtained using the Integrated Gradients algorithm (Sundararajan et al., 2017), to their mean attention weights. I computed the Pearson's correlation coefficient (PCC) between the mean attention weights of fine-tuned BERT of all heads across the last layers (9-12) and the tokens' absolute importance score, as it has been shown that fine-tuning effects mostly BERT's last layers (9-12) (Rogers et al., 2021).

The results show that even though the patterns of the attention weights of fine-tuned BERT are different from those of BERT without fine-tuning, results show that attention weights are not meaningful when it comes to the model's prediction. As I found no linear correlation between the absolute importance score and the mean attention weights of BERT, Table 3, for the examined datasets ($0.056 \leq$ PCC $\leq 0.171$), as well as between the number of occurrences of a token and the mean attention weights ($-0.101 \leq$ PCC $\leq -0.015$) or the mean importance scores ($-0.011 \leq$ PCC $\leq -0.002$). These results suggest that attention weights don't play a direct role in explaining BERT's performance, which is in line with previous studies (Sun and Lu, 2020; Serrano and Smith, 2019; Vashishth et al., 2019).

Finally, I analyzed the importance scores of POS tags of fine-tuned BERT to find out the features that BERT relies on to make its prediction. The results show that BERT captures syntactical biases in the datasets. As the results in Figure 1 show that the POS tags with the highest importance scores are auxiliaries, punctuation, determiners, adpositions, and pronouns which are not informative for the task of hate speech and abuse detection. Among these, the most informative tag for hate speech and abuse detection is the pronoun. These results suggest that BERT relies on syntactical biases and shortcuts in the datasets for its good performance. I

| 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 4: NOI words and the group they describe.

speculate that this syntactical bias is resulted from the upstream datasets that BERT was pre-trained on. To mitigate the effect of that bias, I fine-tuned BERT on an intermediate task which is English POS tags classification dataset following the work suggested in (Zhou et al., 2020). However the results show almost the same distribution of the feature importance scores. This results suggest that Post-processing bias mitigation in BERT is not effective and mitigating the bias during the pre-training might be more effective. The results in this section motivate the second and the third research objectives.

### 3.2 Research objective 2

To achieve my second research goal and to find out if there are other biases in the commonly used word embeddings that are used in the task of hate speech and abuse detection models, I aim to reveal whether word embeddings associate offensive language with words describing marginalized groups. I define systematic offensive stereotypes (SOS) from a statistical perspective as "*A systematic association in the word embeddings between profanity and marginalized groups of people*". In other words, SOS refers to associating offensive terms to different groups of people, especially marginalized people, based on their ethnicity, gender, or sexual orientation. Based on my definition of SOS, I want a method to measure the association that each word embedding model has between profanity and marginalized groups of people. I propose to measure that association using the cosine similarity between swear words and words that describe marginalized social groups.

For the swear words, I used a list of 427 swear words from (Agrawal and Awekar, 2018b). For

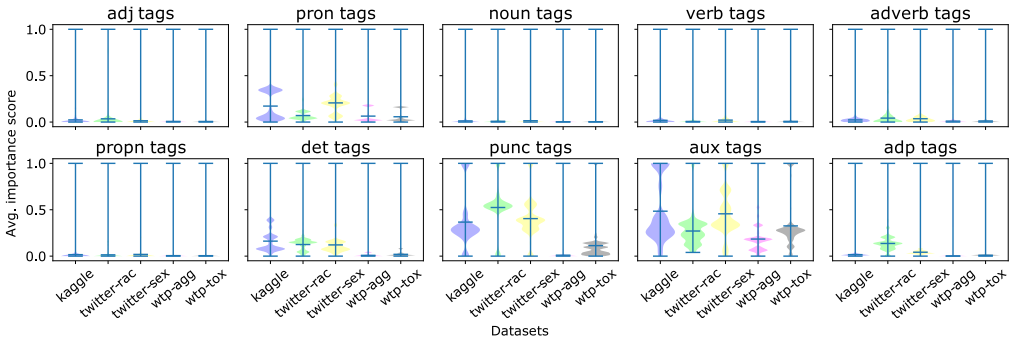


Figure 1: Mean normalised importance scores assigned by fine-tuned BERT to POS tags in the datasets.

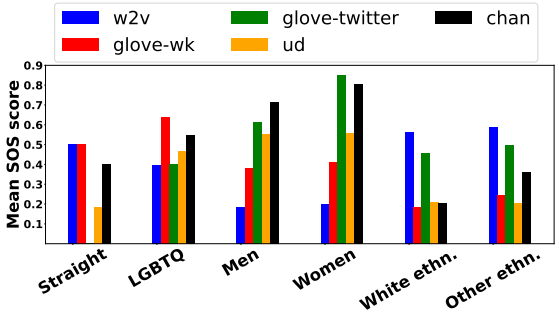


Figure 2: Mean SOS scores for the examined word embeddings and groups.

describing marginalized social groups, I used a word list that contains non-offensive identity (NOI) names to describe marginalized groups of people (Zhou et al., 2021; Dixon et al., 2018) and non-marginalized ones (Sweeney and Najafian, 2019), as summarised in Table 4. Similar to RNSB, I 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). Moreover, According to the reported coherence scores in (Antoniak and Mimno, 2021), The used NOI words for women, men, white and non-white ethnicity groups, score the highest coherence which are 0.090 and 0.910 respectively which shows that the NOI that describe two different groups, e.g. Women vs Men, are far apart which is ideal. However, they don't provide analysis for seed words related to sexual orientation. Since we used the same method to collect these seed words like gender and ethnicity related seed words, I assume that sexual oriented seed words would also have accepted coherence scores.

To measure the SOS bias, 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 $we$, I 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 ($W_{NOI}$).

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

The normalized SOS score takes values within the range $[0, 1]$ and indicates the similarity of an 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 comparatively 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.

I computed the mean SOS score over the examined word embeddings(Word2Vec, Glove-WK, Glove-Twitter, UD, and Chan) for each examined group individually. Figure 2 shows that some word embeddings are more biased than others and that the biased word embeddings are more biased towards the marginalized group than the non-marginalized groups.

To validate my SOS bias metric, I compared the SOS bias, measured by my proposed method and state-of-the-art metrics (WEAT, RNSB, RND, ECT), to the published statistics on online abuse

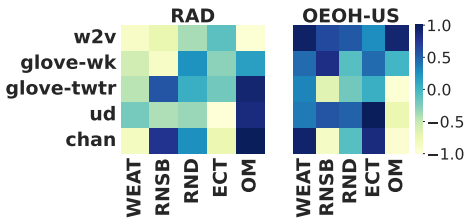


Figure 3: ThePearson's correlation between the different metrics and the percentages of people belonging to the examined marginalized groups who experienced abuse and extremism online for each published surveys for the examined word embedding. For RAD heatmap, correlation is computed between the SOS scores and the differences in RAD between the percentage of (women and men), (LGTBQ and straight), and (Non-white ethnicities and White ethnicities).

and extremism that is targeted at marginalized 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 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.

To measure the $SOS_{women}$ using the state-of-the-art metrics, target list W1 contained the NOI words that describe women in Table 4, target list W2 contained the NOI words that describe men, attribute list 1 contained the same swear words used earlier to measure the SOS bias, and attribute list 2 a list of positive words provided by the WEFE framework. To measure the $SOS_{ethnicity}$, I 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 $SOS_{lgbtq}$, I 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 my proposed metric, I 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 marginalized groups who experienced abuse and extremism online were then acquired fromthe 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 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, I 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 marginalized groups who experienced abuse and extremism online. The results in Figure 3† show that my proposed SOS bias metric, for Chan, UD, and Glove-Twitter, has a high positive correlation with the published statistics on online abuse (RAD), 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), my 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) 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 my 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 marginalized people, while word embeddings that were trained on Google news and Wikipedia articles encode the hate and extremism against the marginalized 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 I 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 latter were omitted from the figure.

capture that difference between the word embeddings. Consequently, the results suggest that my bias metric is more reflective of the SOS bias in the different word embeddings than the-state-of-the-art bias metrics.

| 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 my 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.

I also investigate the influence that my SOS bias metric and state-of-the-art metrics have on the downstream task of hate speech detection. By correlating the F1 scores of machine learning models on different hate speech datasets (Table 5) and the SOS bias scores as measured by my proposed methods and the state-of-the-art metrics. The results in Table 6 show that my metric exhibits a 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 my SOS bias metric correlates consistently positively with the F1 scores of the deep learning models on the different datasets compared to the other metrics. My findings in this section suggest that there is an influence of the SOS bias in the word embeddings on the downstream task of hate speech detection. However, the results are not conclusive and more experiments are required.

The results in this section suggest that the SOS bias provides important information to be used in addition to the social bias to get a fuller picture of the bias in the word embeddings. They also suggest that impact of the SOS and the social bias in the word embeddings on the performance of hate speech detection models. Which means it is important for the future studies on hate speech detection to pay attention to the influence of bias on the models' performance to develop fairer models.

My findings in this section motivate my next research objective to use counterfactual causal inference to understand the influence of the bias in word embedding on the downstream tasks of hate speech and abuse detection.

### 3.3 Research objective 3

This research goal can be achieved by answering the following research question: 1) How to measure the intersectional bias in pre-trained contextual word embeddings? 2) What is the causal influence of bias, social and intersectional, in the pre-trained contextual word embeddings on the task of hate speech detection? and how harmful that bias is it on the models' fairness?

To answer the first research question and to measure the intersectional bias (gender and race) in contextual word embeddings, I plan to first create an intersectional bias dataset similar to SteroSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020) bias datasets but with focus on intersectionality of gender and race. Then, I plan to use the same method proposed to measure the bias in contextual word embeddings using the same method proposed in (Nangia et al., 2020).

To answer the first part of the second questions and to measure the influence of social and intersectional bias on the task of hate speech detection, I plan to compute the Average Treatment Effect (ATE) on the model's prediction probability distribution (Feder et al., 2021b). I plan to compute the ATE of the prediction probability distribution of a biased contextual word embeddings on a hate speech dataset (factual) and the prediction probability distribution of a debiased contextual word embeddings (counterfactual) on hate speech datasets. I plan to use contextual word embeddings without fine-tuning to avoid unobserved con-founders like the bias in the hate speech datasets.

To answer the second half of the second research question and measure the potential harm of the

bias on the task of hate speech detection, I plan to measure the unfairness model for the marginalized and the non-marginalized groups. To measure the unfairness of hate speech detection models, I plan to use similar fairness metric to the one suggested in (De-Arteaga et al., 2019) where the authors measure the difference of the true positive rates (TPR) scores between the different groups of people (marginalised vs. non-marginalized). But instead of the TPRs scores, I plan to use the false positive rate (FPR) scores since FPR is a better estimate of unfairness in hate speech detection models as suggested by (Dixon et al., 2018). Our metric to measure unfairness in hate speech models is described in Equation 2 where $g$ is the marginalized group of people (women, non-white ethnicities, and LGBTQ) and $\hat{g}$ is hte non-marginalized groups of peaple (men, white-ethnciities, and straight).

$$Unfairness_{g,y} = FPR_g - FPR_{\hat{g}} \quad (2)$$

Similarly I plan to use contextual word embeddings without fine-tuning to avoid the unfairness that might result from the imbalances in the datasets. For the experiments I plan to use distilled versions of different pre-trained contextual word embedding, e.g. Distill-BERT, Distill-Roberta, and Distill-GPT2. due to limited access to computational resources. I also plan to use the hate speech datasets described in Table 7, as they contain detailed information on the target of the hate based on attributes like race, gender, and sexual orientation.

This work is expected to reveal the intersectional bias in the contextual word embeddings and how, in addition to the social bias, it causally influence the performance and the unfairness of the hate speech detection models. Understanding this causal influence on performance and fairness would be helpful in developing more effective and targeted debias techniques that address the unfairness of the hate speech detection models instead of generic superficial debias techniques (Gonen and Goldberg, 2019).

| Dataset | Size | |
|---|---|---|
| ETHOS | 433 | (Mollas et al., 2022) |
| MLMa | 5647 | (Ousidhoum et al., 2019) |
| Jigsaw | 1,902,194 | (Jigsaw, 2019) |
| MIT | 59,179 | (Huang et al., 2020) |
| SBIC | 112,900 | (Sap et al., 2020) |

Table 7: Targeted Hate speech datasets

### 3.4 Limitations

Even though this work has a positive implications , it also has its limitations. One of the limitations is studying bias only from the western society perspective as the way bias is measured might differ in different societies. As for intersectional bias, this work focus only on the intersectionality of gender and race. This work focuses only on models and datasets that are in English which is another limitation. Finally, this work studies the influence of bias only on hate speech detection models using only supervised machine learning models.

### 3.5 Ethical consideration

This work has a positive impact on the society since it is targeted at revealing the different biases in the commonly used NLP models. It gives insight into the potential risks and unfairness of these NLP models.

## 4 Conclusion

Hate speech and abuse detection is a very important task to provide a safe inclusive environment for people from different backgrounds to express themselves. However, the different types of biases that have been shown in different NLP tasks could have a counter effect on these hate speech and abuse detection models as they could associate minorities with hate and abuse which could lead to flagging their content as inappropriate and silencing which is the exact opposite of the aim of hate speech and abuse detection models. In this thesis, I look at the different biases in hate speech and abuse detection models and what is the influence of that bias on the performance of hate speech detection models and how this bias could harm the model's fairness. This work reveal types of biases other than social bias in some of the most common NLP models. And it gives insight into developing targeted and effective techniques to mitigate the effect of the different biases and to develop fairer hate speech detection models.

## 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. 2018a. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval*, pages 141–153, Cham. Springer International Publishing.

Sweta Agrawal and Amit Awekar. 2018b. 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.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139.

Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Advances in Artificial Intelligence*, pages 275–281, Cham. Springer International Publishing.

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.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM.

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.

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

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *5th International AAAI Conference on Weblogs and Social Media*.

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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics*, 9:160–175.

Fatma Elsafoury. BERT attention explanation. https://github.com/efatmae/BERT_Attention_Explanation.

Fatma Elsafoury, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. 2021. When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access*, 9:103541–103563.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021a. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021b. Causalm: Causal model explanation through counterfactual language models. *Comput. Linguistics*, 47(2):333–386.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

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.

Paula Giddings. 2006. *When and where I enter*. Bantam Doubleday Dell Publishing Group Incorporated.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL-HLT*.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

James Hawdon, Atte Oksanen, and Pekka Räsänen. 2015. Online extremism and online hate. *NORDICOM*, page 29.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1440–1448. European Language Resources Association.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. Accessed: 2022-02-15.

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.

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

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *EMNLP/IJCNLP (1)*, pages 4364–4373. Association for Computational Linguistics.

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

Akshi Kumar, Shashwat Nayak, and Navya Chandra. 2019. Empirical analysis of supervised machine learning techniques for cyberbullying detection. In *International Conference on Innovative Computing and Communications*, pages 223–230, Singapore. Springer Singapore.

Michael A. Lepori. 2020. Unequal representations: Analyzing intersectional biases in word embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1720–1728. International Committee on 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.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5356–5371. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Laurie T O'Brien, Alison Blodorn, Glenn Adams, Donna M Garcia, and Elliott Hammer. 2015. Ethnic variation in gender-stem stereotypes and stem participation: An intersectional approach. *Cultural Diversity and Ethnic Minority Psychology*, 21(2):169.

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

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Sayanta Paul and Sriparna Saha. 2020. Cyberbert: Bert for cyberbullying identification. *Multimedia Systems*.

John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th international Workshop on Semantic Evaluation*, pages 571–576.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5434–5445. Association for Computational Linguistics.

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

Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in vine. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, page 617–622. ACM.

E. Raisi and B. Huang. 2017. Cyberbullying detection with weakly supervised machine learning. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 409–416.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 194–209. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sarah Rosenfield. 2012. Triple jeopardy? mental health at the intersection of gender, race, and class. *Social Science & Medicine*, 74(11):1791–1801.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.

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

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328.

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.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13209–13220.

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

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if this modified that? syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online. Association for Computational Linguistics.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *arXiv.org*, arXiv:1909.11218.

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 NAACL Student Research Workshop*, pages 88–93, San Diego, California. 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 Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics.

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

Jaideep Yadav, Devesh Kumar, and Dheeraj Chauhan. 2020. Cyberbullying detection using pre-trained bert model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1096–1100. IEEE.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.

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