Thesis Distillation: Investigating The Impact of Bias in NLP Models on Hate Speech Detection

Fatma Elsafoury
Fraunhofer Research Institute (FOKUS), Berlin, Germany
fatma.elsafoury@fokus.fraunhofer.de

Abstract

This paper is a summary of the work done in my PhD thesis. Where I investigate the impact of bias in NLP models on the task of hate speech detection from three perspectives: explainability, offensive stereotyping bias, and fairness. Then, I discuss the main takeaways from my thesis and how they can benefit the broader NLP community. Finally, I discuss important future research directions. The findings of my thesis suggest that the bias in NLP models impacts the task of hate speech detection from all three perspectives. And that unless we start incorporating social sciences in studying bias in NLP models, we will not effectively overcome the current limitations of measuring and mitigating bias in NLP models.

1 Introduction

Hate speech on social media has severe negative impacts, not only on its victims (Sticca et al., 2013) but also on the moderators of social media platforms (Roberts, 2019). This is why it is crucial to develop tools for automated hate speech detection. These tools should provide a safer environment for individuals, especially for members of marginalized groups, to express themselves online. However, recent research shows that current hate speech detection models falsely flag content written by members of marginalized communities, as hateful (Sap et al., 2019; Dixon et al., 2018; Mchangama et al., 2021). Similarly, recent research indicates that there are social biases in natural language processing (NLP) models (Garg et al., 2018; Nangia et al., 2020; Kurita et al., 2019; Ousidhoum et al., 2021; Nozza et al., 2021, 2022).

Yet, the impact of these biases on the task of hate speech detection has been understudied. In my thesis, I identify and study three research problems: 1) the impact of bias in NLP models on the performance and explainability of hate speech detection models; 2) the impact of the imbalanced representation of hateful content on the bias in NLP models; and 3) the impact of bias in NLP models on the fairness of hate speech detection models.

Investigating and understanding the impact of bias in NLP on hate speech detection models will help the NLP community to develop more reliable, effective, and fair hate speech detection models. My research findings can be extended to the general task of text classification. Similarly, understanding the origins of bias in NLP models and the limitations of the current research on bias and fairness in NLP models, will help the NLP community develop more effective methods to expose and mitigate the bias in NLP models.

In my thesis and this paper, I, first, critically review the literature on hate speech detection (§2) and bias and fairness in NLP models (§3). Then, I address the identified research problems in hate speech detection, by investigating the impact of bias in NLP models on hate speech detection models from three perspectives: 1) the explainability perspective (§4), where I address the first research problem and investigate the impact of bias in NLP models on their performance of hate speech detection and whether the bias in NLP models explains their performance on hate speech detection; 2) the offensive stereotyping bias perspective (§5), where I address the second research problem and investigate the impact of imbalanced representations and co-occurrences of hateful content with marginalized identity groups on the bias of NLP models; and 3) the fairness perspective (§6), where I address the third research problem and investigate the impact of bias in NLP models on the fairness of the task of hate speech detection. For each research problem, I summarize the work done to highlight its main findings, contributions, and limitations. Thereafter, I discuss the general takeaways from my thesis and how it can benefit the NLP community at large (§7). Finally, I present directions for future research (§8).

The findings of my thesis suggest that the bias in NLP models has an impact on hate speech detection models from all three perspectives. This means that we need to mitigate the bias in NLP models so that we can ensure the reliability of hate speech detection models. Additionally, I argue that the limitations and criticisms of the currently used methods to measure and mitigate bias in NLP models are direct results of failing to incorporate relevant literature from social sciences. I build on my findings on hate speech detection and provide a list of actionable recommendations to improve the fairness of the task of text classification as a short time solution. For a long-term solution to mitigate the bias in NLP models, I propose a list of recommendations to address bias in NLP models by addressing the underlying causes of bias from a social science perspective.

2 Survey: Hate speech

In Elsafoury et al. (2021a), I provide a comprehensive literature review on hate speech and its different forms. Furthermore, I review the literature of hate speech detection for different methods proposed in the literature accomplishing every step in the text classification pipeline. Then, I point out the limitations and challenges of the current research on hate speech detection.

The main contributions of this survey are: 1) There are different definitions and forms of hate speech. One of the main limitations of current studies on hate speech detection, is the lack of distinction between hate speech and other concepts like cyberbullying. 2) There are many resources of hate speech related datasets in the literature, that allow the development of new hate speech detection models. However, these datasets have many limitations, including limited languages, biased annotations, class imbalances, and user distribution imbalances. 3) One of the main limitations of the current research on hate speech detection, is the lack of understanding how it is impacted by the bias in NLP models. This limitation is what I aim to address in my thesis.

Limitations: One of the main limitations of this survey, is that it focuses on hate speech detection only as a supervised text classification task. However, recent studies propose a framework to automate and enforce moderation policies, instead of training machine learning models to detect hate speech (Calabrese et al., 2022). Similarly, this review focuses on hate speech datasets that are collected only from social media platforms. However, recently, generative models have become more popular and started to be used in generating hate speech related datasets (Hartvigsen et al., 2022).

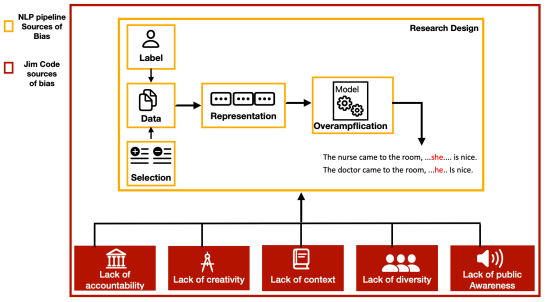

Figure 1: The sources of bias in supervised NLP models

3 Survey: Bias and Fairness in NLP

In Elsafoury and Abercrombie (2023), I review the literature on the definitions of bias and fairness in NLP models. Additionally, I review the literature on the origins of bias in NLP models from two perspectives: 1) NLP pipeline as discussed in Shah et al. (2020); Hovy and Prabhumoye (2021), and 2) social sciences and critical race theory as discussed in Benjamin (2019); Broussard (2023); Nobel (2018).

There are many definitions of the term *bias*. The normative definition of bias, in cognitive science, is: *"Behaving according to some cognitive priors and presumed realities that might not be true at all"* (Garrido-Muñoz et al., 2021). And the statistical definition of bias is "*A systematic distortion in the sampled data that compromises its representatives"* (Olteanu et al., 2019). The statistical definition of bias is the one used in this thesis.

In this work, I argue that the sources of bias in the NLP pipeline originate in the social sciences and that they are direct results of the sources of bias from the social science (Jim code) perspective as shown in Figure 1.

The main contribution of this literature review is reviewing the sources of bias in NLP models from the social science perspective as well as the NLP perspective. This survey points out the limitations of the currently used methods to measure and mitigate bias in NLP models. It also suggests that these limitations are direct results of the lack of inclusion of social science literature in the development of methods that quantify and

mitigate bias in NLP. Finally, I share a list of actionable suggestions and recommendations with the NLP community on how to mitigate the limitations discussed in studying bias in NLP (§7).

Limitations: One main limitation of this survey is that it reviews the literature on the sources of bias in the NLP pipeline, only in supervised models. Unsupervised NLP models might have different sources of bias. The second limitation is regarding the reviewed literature on the sources of bias in social sciences, where I rely mainly on three books *Algorithms of Oppression: How Search Engines Reinforce Racism* by Safiya Nobel (Nobel, 2018), *Race after Technology: Abolitionist Tools for the New Jim Code* by Ruha Benjamin Benjamin (2019), and *More than a glitch: Confronting race, gender, and ability bias in tech* by Meredith Broussard (Broussard, 2023). A more comprehensive literature review to review studies that investigate the direct impact of social causes on bias in NLP would be important future work. However, to the best of my knowledge, this area is currently understudied.

In the next sections, I address the understudied impact of bias in NLP models on hate speech detection models. I investigate that impact from the following perspectives.

4 The explainability perspective

For this perspective, I investigate the performance of different hate speech detection models and whether the bias in NLP models explains their performance on the task of hate speech detection. To achieve that, I investigate two sources of bias:

1. **Bias introduced by pre-training:** where I investigate the role that pre-training a language model has on the model's performance, especially when we don't know the bias in the pre-training dataset. I investigate the explainability of the performance of contextual word embeddings, also known as language models (LMs), on the task of hate speech detection. I analyze BERT's attention weights and BERT's feature importance scores. I also investigate the most important part of speech (POS) tags that BERT relies on for its performance. The results of this work suggest that pre-training BERT results in a syntactical bias that impacts its performance on the task of hate speech detection (Elsafoury et al., 2021b).

Based on these findings, I investigate whether the social bias resulting from pre-training contextual word embeddings explains their performance on hate speech detection in the same way syntactical bias does. I inspect the social bias in three LMs (BERT (base and large) (Devlin et al., 2019), ALBERT (base and xx-large) (Lan et al., 2020), and ROBERTA (base and large) (Liu et al., 2019)) using three different bias metrics, CrowS-Pairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), and SEAT (May et al., 2019), to measure gender, racial and religion biases. First, I investigate whether large models are more socially biased than base models. The Wilcoxon statistical significance test (Zimmerman and Zumbo, 1993) indicates that there is no statistical significant difference between the bias in base and large models in BERT and RoBERTa, unlike the findings of (Nadeem et al., 2021). However, there is a significant difference between the base and xx-large ALBERT. These results suggest that large models are not necessarily more biased than base models, but if the model size gets even bigger, like ALBERT-xx-large, then the models might get significantly more biased. Since there is no significant difference between the base and large models, I only use base LMs in the rest of the thesis.

Then, I follow the work of (Steed et al., 2022; Goldfarb-Tarrant et al., 2021) and use correlation as a measure of the impact of bias on the performance of the task of hate speech detection. The Pearson's correlation coefficients between the bias scores of the different models and the F1-scores of the different models on the used five hate-speech-related datasets are inconsistently positive as shown in Figure 2. However, due to the limitations of the metric used to measure social bias, as explained in Blodgett et al. (2021), the impact of the social bias in contextual word embeddings on their performance on the task of hate speech detection remains inconclusive.

2. **Bias in pre-training datasets:** Where I investigate the impact of using NLP models pre-trained on data collected from social media platforms like Urban dictionary and 4 & 8 Chan, which are famous for having sexist and racist posts (Nguyen et al., 2017; Papasavva et al., 2020). I investigate the performance of two groups of static word embeddings (SWE) on hate speech detection. The first group, social-media-based, pre-trained on biased datasets that contain hateful content. This group consists of Glove-Twitter (Mozafari

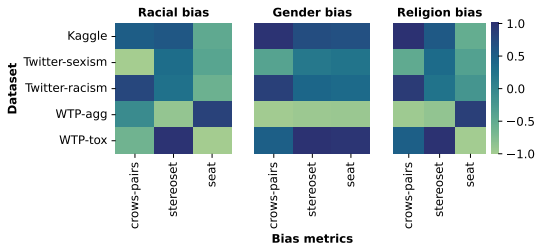

Figure 2: Heatmap of the Pearson correlation coefficients between the performance (F1-scores) of LMS on the different hate speech datasets and the social bias scores.

et al., 2020), Urban dictionary (UD) (Wilson et al., 2020), and 4& 8 Chan (chan) (Voué et al., 2020) word embeddings. The second group of word embeddings, informational-based, is pre-trained on informational data collected from Wikipedia and Google New platforms. This group contains the word2vec (Mikolov et al., 2021) and Glove-WK word (Pennington et al., 2014) embeddings. SWE in this part of the work because there are SWE that are pre-trained on datasets collected from social media platforms like urban dictionary, and 4 &8 Chan. First, I investigate the ability of the five different word embeddings, to categorize offensive terms in the Hurtlex lexicon. Then, I investigate the performance of Bi-LSTM model with an un-trainable embeddings layer of the five word embeddings on the used five hate-speech-related datasets. The results indicate that the word embeddings that are pre-trained on biased datasets social-media-based, outperform the other word embeddings that are trained on informational data, informational-based on the tasks of offenses categorization and hate speech detection (Elsafoury et al., 2022b).

Based on these findings, I inspect the impact of social bias, gender, and racial, in the SWE on their performance on the task of hate speech detection. To measure the social bias in the SWE, I use the following metrics from the literature: WEAT (Caliskan et al., 2017), RNSB (Sweeney and Najafian, 2019), RND (Garg et al., 2018), and ECT (Dev and Phillips, 2019). Then, I use Pearson's correlation to investigate whether the social bias in the word embeddings explains their performance on the task of hate speech detection. Similar to LMs, the results indicate an inconsistent positive correlation between the bias scores and the F1-sores of the Bi-LSTM model using the different word embeddings as shown in Figure 3. This lack of positive correlation could be due to limitations in the used metrics to measure social bias in SWE (Antoniak and Mimno, 2021). These results suggest that the impact of the social bias in the SWE on the performance of the task of hate speech detection is inconclusive.

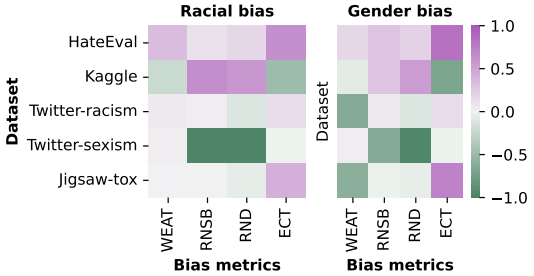

Figure 3: Heatmap of the Pearson correlation coefficients between the performance (F1-scores) of SWE on the different hate speech datasets and the social bias scores.

Contributions: The main findings and contributions of the explainability perspective can be summarized as: **1)** The results provide evidence that the syntactical bias in contextual word embeddings, resulting from pre-training, explains their performance on the task of hate speech detection. **2)** The results suggest that pre-training static word embeddings on biased datasets from social-media-based sources improves and might explain the performance of the word embeddings on the task of hate speech detection. **3)** For both static and contextual word embeddings, there is no strong evidence that social bias explains the performance of hate speech detection models. However, due to the limitations of the methods used to measure social bias in both static and contextual word embeddings, this finding remains inconclusive.

Limitations: one of the main limitations of this work is using social bias metrics from the literature, which have their limitations as argued in Blodgett et al. (2021); Antoniak and Mimno (2021). Additionally, the work done here, is limited to hate speech datasets that are in English. Similarly, the social bias inspected in the different word embeddings is based on Western societies, where the marginalized groups might be different in different societies. It is also important to mention that the findings of this work are limited to the used datasets and models and might not generalize to other models or datasets.

5 The offensive stereotyping bias perspective

In Elsafoury et al. (2022a); Elsafoury (2023), I investigate how the hateful content on social media and other platforms that are used to collect data and pre-train NLP models, is being encoded by those NLP models to form systematic offensive stereotyping (SOS) bias against marginalized groups of people. Especially with imbalanced representation and co-occurrence of the hateful content with the marginalized identity groups. I introduce the systematic offensive stereotyping (SOS) bias and formally define it as "*A systematic association in the word embeddings between profanity and marginalized groups of people.*" (Elsafoury, 2022).

I propose a method to measure it and validate it in static (Elsafoury et al., 2022a) and contextual word embeddings (Elsafoury et al., 2022a). Finally, I study how it impacts the performance of these word embeddings on hate speech detection models. I propose the normalized cosine similarity to profanity (NCSP) metric, which is a metric to measure the SOS bias in static word embeddings using the cosine similarity between a list of swear words and non-offensive identity (NOI) words that describe three marginalized groups (Women, LGBTQ, and Non-White) described in Table 1. As for measuring the SOS bias in contextual word embeddings, I propose the SOS_{LM} metric. The SOS_{LM} metric uses the masked language model (MLM) task to measure the SOS bias, similar to the work proposed in StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020) metrics. Instead of using crowdsourced sentence pairs that express socially biased sentences and socially unbiased sentences, I use synthesized sentence pairs that express profane sentences and non-profane sentence-pairs. I measure the SOS bias scores in 15 static word embeddings (Elsafoury et al., 2022a) and 3 contextual word embeddings (Elsafoury, 2023). The results show that for static word embeddings, there is SOS bias in all the inspected word embeddings, and it is significantly higher towards marginalized groups as shown in table 2. Similarly, Figure 4 show that all the inspected contextual word embeddings are SOS biased, but the SOS bias scores are not always higher towards marginalized groups. Then, I validate the SOS bias itself by investigating how reflective it is of the hate that the same marginalized groups experience online. The correlation results, using Pearson correlation coefficient, indicate that there is a positive correlation between the measured SOS bias in static and contextual word embeddings and the published statistics of the percentages of the marginalized groups (Women, LGBTQ, and non-white ethnicities) that experience online hate (Hawdon et al., 2015) and the measured SOS bias scores in static word embeddings using the NCSP metric and the SOS_{LM} metric. I also validate the proposed metric to measure the SOS bias in comparison to the social bias metrics proposed in the literature. I use the Pearson correlation coefficient between the social bias scores and the SOS bias scores in the static and the contextual

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

Table 1: The non-offensive identity (NOI) words used to describe the marginalized and non-marginalized groups in each sensitive attribute. For the disability-sensitive attributes, we use only words to describe disability due to the lack of words used to describe able-bodied.

Word embeddings	**Mean SOS**		
	Women	LGBTQ	Non-white
W2V	0.293	**0.475**	0.456
Glove-WK	0.435	**0.669**	0.234
glove-twitter	**0.679**	0.454	0.464
UD	0.509	**0.582**	0.282
Chan	**0.880**	0.616	0.326
Glove-CC	**0.567**	0.480	0.446
Glove-CC-large	0.318	0.472	**0.548**
FT-CC	0.284	**0.503**	0.494
FT-CC-sws	0.473	0.445	**0.531**
FT-WK	0.528	**0.555**	0.393
FT-WK-sws	**0.684**	0.656	0.555
SSWE	0.619	0.438	**0.688**
Debias-W2V	0.205	0.446	**0.471**
P-DeSIP	0.266	**0.615**	0.354
U-DeSIP	0.266	**0.616**	0.343

Table 2: The mean SOS bias score of each static word embeddings towards each marginalized group. Bold scores reflect the group that the static word embeddings is most biased against (Elsafoury et al., 2022a).

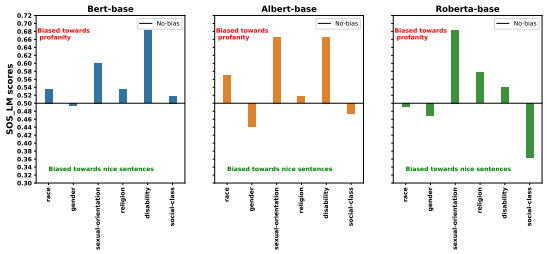

Figure 4: SOS_{LM} bias scores in the different language models (Elsafoury, 2023).

word embeddings. The results show that, for the inspected static word embeddings, the correlation results, according to Pearson correlation, show a negative correlation between the measured SOS bias scores measured using the NCSP metric and the social bias scores (gender and race) measured using the WEAT, RND, RNSB, and ECT metrics. As for the contextual word embeddings, the Pearson correlation coefficient results show a positive correlation between the SOS bias scores measured using the SOS_{LM} metric and the social bias scores (gender, race, and religion) measured using the CrowS-Pairs metric, which could be the case because the SOS_{LM} metric is built on the CrowS-Pairs metric.

Finally, I investigate whether the inspected SOS bias explained the performance of the inspected word embeddings on the task of hate speech detection. I train MLP and Bi-LSTM models with an untrainable layer of the different static word embeddings on four hate-speech-related datasets. As for contextual word embeddings, I fine-tune BERT-base-uncased, ALBERT-base, and ROBERTA-base on six hate speech related datasets. Then, I use Pearson's correlation between the SOS bias scores in the different word embeddings and their F1 scores on the models on the task of hate speech detection. The correlation results, similar to the results in §4, show an inconsistent positive correlation. This could be because the limitations of other social bias metrics in the literature are extended to the proposed metrics. In this case, the impact of the SOS bias in static and contextual word embeddings on their performance on the task of hate speech detection remains inconclusive.

Contributions: The main findings and contributions of the offensive stereotyping perspective can be summarized as follows: **1)** I define the SOS bias, propose two metrics to measure it in static and contextual word embeddings, and demonstrate that SOS bias correlates positively with the hate that marginalized people experience online. **2)** The results of this section provide evidence that all the examined static and contextual word embeddings are SOS biased. This SOS bias is significantly higher for marginalized groups in static word embeddings versus non-marginalized groups. However, this is not the case with the contextual word embeddings. **3)** Similar to social bias, there is no strong evidence that the SOS bias explains the performance of the different word embeddings on the task of hate speech detection.

Limitations: The findings of this work are limited to the examined word embeddings, models, and datasets, and might not generalize to others. Similarly, the SOS bias scores measured using the NCSP metric in the inspected static word embeddings, are limited to the used word lists. Another limitation is regarding my definition of the SOS bias, as I define bias from a statistical perspective, which lacks the social science perspective as discussed in Blodgett et al. (2021); Delobelle et al. (2022). Moreover, I only study bias in Western societies where Women, LGBTQ and Non-White ethnicities are among the marginalized groups. However, marginalized groups could include different groups of people in other societies. I also only use datasets and word lists in English, which limits our study to the English-speaking world. Similar to other works on quantifying bias, our proposed metric measures the existence of bias and not its absence (May et al., 2019), and thus low bias scores do not necessarily mean the absence of bias or discrimination in the word embeddings. Another limitation of this work is the use of template sentence-pairs to measure the SOS bias in contextual word embeddings, which do not provide a real context that might have impacted the measured SOS bias. Since the proposed method used to measure the SOS bias in contextual word embeddings (SOS_{LM}) builds on social bias metrics like CrowS-Pairs and StereoSet, it is highly likely that SOS_{LM} have the same limitations as CrowS-Pairs and StereoSet that are pointed out in Blodgett et al. (2021).

6 The fairness perspective

In Elsafoury et al. (2023), I investigate how different sources of bias in NLP models and their removal impact the fairness of the task of hate

speech detection. Improving the fairness of the text classification task is very critical to ensure that the decisions made by the models are not based on sensitive attributes like race or gender.

I first measure three sources of bias according to (Shah et al., 2020; Hovy and Prabhumoye, 2021): representation bias, selection bias, and overamplification bias. Then, I fine-tune three language models: BERT, ALBERT, and ROBERTA on the Jigsaw dataset (Jigsaw, 2018), and measure the fairness of these models using two sets of fairness metrics: threshold-based and threshold-agnostic. The threshold-based metrics are the TPR_gap and the FPR_gap metrics used in Steed et al. (2022); De-Arteaga et al. (2019). As for the threshold-agnostic metric, I use the AUC_gap metric, which is an adaptation of the metrics proposed in Borkan et al. (2019). I investigate the impact of the different sources of bias on the models' fairness by measuring the Pearson correlation coefficient between the bias scores and the fairness score. Then, I investigate the impact of removing the three sources of bias, using different debiasing methods, on the fairness of hate speech detection models. I remove the representation bias using the SentDebias method proposed in Liang et al. (2020) to remove gender, racial, religious and SOS bias on the inspected language models. To remove the selection bias, I aim to balance the ratio of positive examples between the identity groups in the Jigsaw dataset. To achieve that, I generate synthetic positive examples using existing positive examples in the Jigsaw training dataset, but with word substitutions using the NLPAUG tool that uses contextual word embeddings to generate word substitutions (Ma, 2019). To remove the overamplification bias, I aim to ensure that the different identity groups, in the Jigsaw dataset, appear in similar semantic contexts in the training dataset, as proposed in Webster et al. (2020). To achieve that, I use different methods: 1) create data perturbations, 2) I use the sentDebias method to remove the bias representations from the fine-tuned models. Thereafter, I compare the fairness of the inspected language models on the task of hate speech detection before and after removing each of the inspected source of bias. I aim to find the most impactful source of bias on the fairness of the task of hate speech detection and to find out the most effective debiasing method. The results suggest that overamplification and selection bias are the most impactful on the fairness of the task of hate speech detection and removing it using data perturbations is the most effective debiasing method. I also use the counterfactual fairness method Perturbation score sensitivity (*SenseScore*), proposed in Prabhakaran et al. (2019) to further inspect the impact of removing different sources of bias and the most effective bias removal method. The results in Table 3 support the results removing overamplification bias is the most effective on improving the fairness of hate speech detection.

	SenseScore		
Model	Gender	Race	Religion
ALBERT-base	$6.9e^{-05}$	0.032	0.006
+ downstream-perturbed-data	↓ $4.2e^{-05}$	↓ 0.002	↓ 0.001
+ downstream-stratified-data	↑ 0.042	0.032	↑ 0.009
+ downstream- stratified-perturbed-data	↑ 0.013	↓ 0.003	↓ 0.0007
BERT-base	0.001	0.03	0.001
+ downstream-perturbed-data	↓ 0.0007	↓ 0.003	0.001
+ downstream-stratified-data	↑ 0.025	↓ 0.022	↑ 0.004
+ downstream- stratified-perturbed-data	↑ 0.002	↓ 0.002	↓ 0.0008
RoBERTa-base	0.001	0.024	0.003
+ downstream-perturbed-data	↓ 0.0008	↓ 0.006	↓ 0.001
+ downstream-stratified-data	↑ 0.038	↑ 0.036	0.003
+ downstream- stratified-perturbed-data	↑ 0.003	↓ 0.002	↓ 0.0003

Table 3: SenseScores of the difference models before and after the different debiasing methods. (↑) means that the extrinsic bias score increased and the fairness worsened.(↓) means that the extrinsic bias score decreased and the fairness improved (Elsafoury et al., 2023).

Finally, I build on the findings of this work and propose practical guidelines to ensure the fairness of the task of text classification and showcase these recommendations on the task of sentiment analysis.

Contributions: The main findings and contributions of the fairness perspective can be summarized as follows: **1)** The results demonstrate that the dataset used to measure the models' fairness on the downstream task of hate speech detection plays an important role in the measured fairness scores. **2)** The results indicate that it is important to have a fairness dataset with similar semantic contexts and ratios of positive examples between the identity groups within the same sensitive attribute, to make sure that the fairness scores are reliable. **3)** Unlike the findings of previous research (Cao et al., 2022; Kaneko et al., 2022), the results demonstrate that there is a positive correlation between representation bias, measured by the CrowS-Pairs and the SOS_{LM} metrics, and the fairness scores of the different models on the downstream task of hate speech detection. **4)** Similar to findings from previous research, (Steed et al., 2022), the results of this work demonstrate that downstream

sources of bias, overamplification and selection, are more impactful than upstream sources of bias, representation bias. **5)** The results also demonstrate that removing overamplification bias by training language models on a dataset with a balanced contextual representation and similar ratios of positive examples between different identity groups, improved the models' fairness consistently across the sensitive attributes and the different fairness metrics, without sacrificing the performance. **6)** I provide empirical guidelines to ensure the fairness of the text classification.

Limitations: It is important to point out that the work done in this section is limited to the examined models and datasets. This work studies bias and fairness from a Western perspective regarding language (English) and culture. There are also issues regarding the datasets that those metrics used to measure the bias (Blodgett et al., 2021). The used fairness metric, extrinsic bias metrics, also received criticism (Hedden, 2021). This means that even though I used more than one metric and different methods to ensure that our findings are reliable, the results could be different when applied to a different dataset. It is also important to mention that there is a possibility that the findings regarding the most effective debiasing method, which is fine-tuning the models on a perturbed dataset, is the case because I use a perturbed fairness dataset as well. I recognize that the provided recommendations to have a fairer text classification task rely on creating perturbations for the training and the fairness dataset. It might be challenging for some datasets, especially if the mention of the different identities is not explicit, like using the word "Asian" to refer to an Asian person but using Asian names instead. Additionally, for the sentiment analysis task, the used keyword to filter the IMDB dataset and get only gendered sentences might provide additional limitations that might have influenced the results. Moreover, in this section, I aim to achieve equity in the fairness of the task of text classification between the different identity groups. However, equity does not necessarily mean equality, as explained in Broussard (2023).

7 What have we learned?

In this section, I combine all the findings of my thesis and point out how this work can benefit the NLP community and the ongoing research on hate speech detection, bias, and fairness in NLP. The survey of the literature on hate speech detection in §2 shows a lack of research on the impact of bias in NLP models and hate speech detection models. Especially the impact on the performance of hate speech detection, and how the hateful content led NLP models to form an offensive stereotyping bias, in addition to limitations with the current research that investigates the impact of bias in NLP models on the fairness of hate speech detection models. The aim of my thesis is to fill these research gaps.

The research goal of my thesis is to investigate the bias in NLP models and its impact on the performance and fairness of the task of hate speech detection, and more generally, the task of text classification. The findings of my thesis show that the bias in NLP models is preventing us from having reliable and effective hate speech detection and text classification models. This is evident by the findings of my thesis.

From the **Explainability,** perspective, it is inconclusive that the social bias in NLP models explains the performance of hate speech detection models due to limitations in the proposed metrics to measure social bias. However, the results in §4 also indicate that the bias resulting from pre-training language models, e.g., syntactic bias and biased pre-training datasets, impacts and explains their performance on hate speech detection modes. This good performance suggests that the hate speech detection model associates hateful content with marginalized groups. This might result in falsely flagging content written by marginalized groups on social media platforms.

From the **Offensive stereotyping bias** perspective, the findings in §5 demonstrate that word embeddings, static and contextual, are systematic offensive stereotyping (SOS) biased. The results show no strong evidence that the SOS bias explains the performance of the word embeddings on the task of hate speech detection, due to limitations in the proposed metrics to measure the SOS bias. However, the existence of SOS bias might have an impact on the hate speech detection models in ways that we have not explored or understood yet, especially against the marginalized groups.

From the **Fairness** perspective, the findings of §6 show that the inspected types of bias, representation, selection, overamplification, have an impact on the fairness of the models on the task of hate speech detection, especially the

downstream sources of bias which are selection and overamplification bias. This means that the bias in the current hate speech datasets and the bias in the most commonly used language models have a negative impact on the fairness of hate speech detection models. Hence, researchers should pay attention to these biases and aim to mitigate them before implementing hate speech detection models.

These findings assert the notion that bias in NLP models negatively impacts hate speech detection models and that, as a community, we need to mitigate those biases so that we can ensure the reliability of hate speech detection models. However, in §3, I discuss the limitations and criticisms of the currently used methods to measure and mitigate bias in NLP models that fail to incorporate findings from the social sciences.

As a short-term solution to improve the fairness of hate speech detection and text classification tasks, I provide a list of guidelines in Elsafoury et al. (2023). These guidelines can be summarized as follows:

1. Measure the bias in the downstream task.
2. Remove overamplification bias.
3. To reliably measure fairness, use a balanced fairness dataset and counterfactual fairness metrics.
4. Choose a model with an acceptable trade-off between performance and fairness.

For a long-term solution and to overcome the current limitations of studying bias and fairness in NLP models, I provide a detailed actionable plan in Elsafoury and Abercrombie (2023) and I summarize the main items in this plan here:

1. Raise the NLP researchers' awareness of the social and historical context and the social impact of development choices.
2. Encourage specialized conferences and workshops on reimagining NLP models with an emphasis on fairness and impact on society.
3. Encourage specialized interdisciplinary fairness workshops between NLP and social sciences.
4. Encourage diversity in NLP research teams.
5. Incorporating more diversity workshops in NLP conferences.
6. Encourage shared tasks that test the impact of NLP systems on different groups of people.

8 Future work

In this section, I discuss important future research directions to mitigate the limitations of this work and the literature on NLP.

8.1 Widening the study of bias in NLP

One of the main limitations of this work and most of the work on bias and fairness in NLP models is that it focuses on the English language and on bias from a Western perspective. A critical future work is to create biased datasets in different languages to investigate social bias in models that are pre-trained on data in different languages. It is also important to investigate bias in multilingual NLP models and bias against marginalized groups in societies apart from Western societies.

8.2 Investigate the impact of social bias causes on the bias in NLP

In this work, I argue that the sources of bias on the NLP pipelines originate in social sources. I also argue that the methods proposed to measure and mitigate bias in NLP models are inefficient, as a result of failing to incorporate social sciences literature and methods. One of the main limitations of this work is the lack of studies that empirically support this argument. This research direction is an important step towards understanding the bias and fairness in NLP and machine learning models in general.

8.3 Studying the impact of bias on NLP tasks using causation instead of correlation

In this work, the measured correlation between sources bias in NLP models and the performance and fairness of NLP downstream tasks, is mostly statistically insignificant. Using causation instead of correlation to investigate that impact could be more effective.

9 Conclusion

In this paper, I provide a summary of my PhD thesis. I describe the work done to each my research findings and contributions. I also discuss the limitations of my work and how they can be mitigated in future research. Moreover, I discuss the main lessons learned from my research as well as recommendations that can benefit the NLP research community, especially for studying and mitigating bias in NP models and improving the fairness of text classification tasks.

References

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.

Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the new jim code*. Polity.

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

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

Meredith Broussard. 2023. *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022. Explainable abuse detection as intent classification and slot filling. *Trans. Assoc. Comput. Linguistics*, 10:1440–1454.

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

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

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*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

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

Sunipa Dev and Jeff M. 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 *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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Fatma Elsafoury. 2023. Systematic offensive stereotyping (sos) bias in language models. *arXiv preprint arXiv:2308.10684*.

Fatma Elsafoury and Gavin Abercrombie. 2023. On the origins of bias in nlp through the lens of the jim code. *arXiv preprint arXiv:2305.09281*.

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

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

Fatma Elsafoury, Stamos Katsigiannis, Steven R. Wilson, and Naeem Ramzan. 2021b. Does BERT pay attention to cyberbullying? In *Proceedings of*

the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, page 1900–1904, New York, NY, USA. Association for Computing Machinery.

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

Fatma Elsafoury, Steven R. Wilson, and Naeem Ramzan. 2022b. A comparative study on word embeddings and social NLP tasks. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Seattle, Washington. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

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

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

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

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

Brian Hedden. 2021. On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs*, 49(2):209–231.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

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

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

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

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

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

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 622–628. Association for Computational Linguistics.

Jacob Mchangama, Natalie Alkiviadou, and Raghav Mendiratta. 2021. A FRAMEWORK OF FIRST REFERENCE Decoding a human rights approach to content moderation in the era of platformization. *The Future of free speech*.

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

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications*

VIII, pages 928–940, Cham. Springer International Publishing.

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

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

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.

Safiya Umoja Nobel. 2018. *Algorithms of Oppression: How search engines reinforce racism*. New York University Press.

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

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

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

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

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, Hong Kong, China. Association for Computational Linguistics.

Sarah Roberts. 2019. *Behind the screens: content moderation in the shadows of social media*. Yale University Press.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

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

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Fabio Sticca, Sabrina Ruggieri, Françoise Alsaker, and Sonja Perren. 2013. Longitudinal risk factors for cyberbullying in adolescence. *Journal of community & applied social psychology*, 23(1):52–67.

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.

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

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and

Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

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.

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.