Debiasing Static Embeddings for Hate Speech Detection

Ling Sun, Soyoung Kim, Xiao Dong, Sandra Kübler
Indiana University
{ls44,sk35,dong1,skuebler}@iu.edu

Abstract

We examine how embedding bias affects hate speech detection by evaluating two debiasing methods—hard-debiasing and soft-debiasing. We analyze stereotype and sentiment associations within the embedding space and assess whether debiased models reduce censorship of marginalized authors while improving detection of hate speech targeting these groups. Our findings highlight how embedding bias propagates into downstream tasks and demonstrates how well different embedding bias metrics can predict bias in hate speech detection.

1 Introduction

Bias in hate speech detection is known to arise from data sources, sampling methods, and pre-trained word embeddings. These different biases distort model predictions, potentially unintentionally linking non-discriminatory terms to hate speech. For instance, Wiegand et al. (2019) showed that domain-restricted sampling methods statistically induce bias, such as the word *commentator* becoming indicative of hate speech because of the content domain *soccer*. Beyond dataset biases, pre-trained word embeddings can encode and amplify historical and social biases from large-scale text data. Bolukbasi et al. (2016) showed that embeddings reinforce stereotypes, such as aligning *man* with *scientist* and *woman* with *homemaker*. Similarly, Caliskan et al. (2017) reported that identity-related terms, such as African American names, are more strongly associated with negative sentiment than European American names.

Such biases undermine hate speech detection by (1) damaging model performance in realistic settings where speech does not conform to learned biases (Wiegand et al., 2019), (2) disproportionately flagging non-hate posts by marginalized groups, reinforcing discrimination, and (3) failing to recognize harmful stereotypes, leading to missed detection of implicit hate against marginalized groups. We will refer to the disproportionate flagging of non-hate posts by marginalized groups as *author bias*; and we will refer to the failure to recognize harmful stereotypes as *target bias*.

While many studies have addressed biases introduced by datasets and sampling strategies (Dixon et al., 2018; Wiegand et al., 2019; Razo and Kübler, 2020), the impact of pre-trained word embeddings on hate speech detection remains underexplored. Not only can the pre-trained embeddings encode inaccurate connotations, they can also reinforce stereotypes which are crucial for detecting implicit hate. Furthermore, Fersini et al. (2023) demonstrated that common debiasing methods for embeddings can introduce new biases and mitigating negative connotation bias may inadvertently reinforce stereotypes. They argue that evaluating debiasing techniques requires assessing the impact on both embedding space associations and downstream task performance. However, in hate speech detection, this dual evaluation remains largely unaddressed.

In this study, we examine two popular debiasing methods, hard-debiasing and soft-debiasing, in the context of hate speech detection. We analyze how these methods alter sentiment and stereotype associations of identity terms within the embedding space, and we evaluate whether debiased models exhibit less bias in hate speech classification. Specifically, we test whether models disproportionately censor authors from marginalized groups and whether they fail to detect hate speech targeting these groups.[1] Our study shows how embedding bias propagates into downstream consequences and evaluates the effectiveness of different embedding bias metrics in predicting bias in hate speech detection.

[1]Our code is available at https://github.com/LingSyrina/hate_speech_bias; our debiased embeddings can be found in https://huggingface.co/datasets/LingSyrina/debiased_embedding

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 67–76

Offensive Content Warning: This report contains some examples of hateful content. This is strictly for the purposes of enabling this research, and we have sought to minimize the number of examples where possible. Please be aware that this content could be offensive and cause you distress.

2 Related Work

2.1 Debiasing Pre-trained Embeddings

Several previous studies suggested debiasing embeddings as potential methods to reduce bias, especially with data augmentation (*e.g.*, Bolukbasi et al., 2016; Park et al., 2018). In contrast, debiased embeddings without data augmentation showed mixed results in that hard-debiasing improves performance for Turkish while decreasing it for English (Şahinuç et al., 2023). However, all of these findings point towards a recurring trend: debiased embeddings can reduce bias in one context but may reduce it in another.

2.2 Bias Metrics and Inconsistency

The inconsistencies in the results discussed above remain unclear, largely because most studies adopted bias metrics that focus on classification performance rather than the embedding space itself. Common bias metrics in hate speech detection, such as False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) (Dixon et al., 2018) assess bias primarily through classification outcomes per identity group. Without directly analyzing embedding space as altered by debiasing, it is difficult to understand why these methods yield mixed results.

Each debiasing method seems to have a distinctive affect on different types of bias. Fersini, Candelier and Pastore (2023) showed that hard and soft debiasing methods can reduce bias in one area (e.g., coherence) while increasing it in another (e.g., stereotype associations). As such, no single debiasing method can address all biases, and selecting the right metrics may align better with specific applications such as hate speech detection. It is thus important to report these different metrics to obtain a comprehensive view.

3 Methodology

Here, we discuss the two datasets (Section 3.1) and the deep learning model (Section 3.3). Following our methodology framework (see Fig. 1), we apply hard and soft debiasing separately to the pre-trained embeddings (Section 3.4) and retrain the GRU models on both datasets. We then assess bias in the embedding space (Section 3.5.1) and in hate speech detection (Section 3.5.2), before and after debiasing. Finally, we test whether embedding bias metrics can predict author and target bias in hate speech detection.

3.1 Hate Speech Datasets

We use (1) the English dataset (MTC-E) by Huang et al. (2020) and (2) the Social Bias Inference Corpus (SBIC) by Sap et al. (2020). Both datasets are annotated for race and gender (among other categories), but MTC-E provides this information about authors while SBIC provides information about that targets of offense.

MTC-E includes 83,077 English tweets, with 36.86% labeled as hate speech and 63.14% as non-hate. Annotations include author demographics: 50.1% male, 49.9% female; 50.5% white, 49.5% non-white.

SBIC consists of 44,671 English tweets annotated for offensiveness. We use only "not offensive" (44.06%) and "offensive" (55.94%) labels for our work. The dataset is annotated for hate targets with multi-class labels. We focus on gender and race annotations, including 24,975 tweets targeting women, 3,615 targeting men, 660 targeting White, 38,880 targeting Black, and 2,850 targeting Asian.[2]

3.2 Embeddings

We use 3,300-dimensional pre-trained embeddings:

Skipgrams from Word2Vec[3] are trained on a portion of the Google News dataset (about 100B words). The model contains 3M words, making it the largest among the 3 embeddings.

FastText comprises 1M word vectors trained on Wikipedia 2017, UMBC web-based corpus and `statmt.org` news[4]. The content is mainly news based, along with other web contents such as blogs.

GloVe is trained on 2B tweets with a 1.2M vocabulary[5]. Unlike the other two news-based em-

[2]Binary gender and three racial terms (White vs. Black vs. Asian) were selected for our study to match the debiasing corpus adopted from Manzini et al. (2019).

[3]https://huggingface.co/fse/word2vec-google-news-300

[4]https://huggingface.co/fse/fasttext-wiki-news-subwords-300

[5]https://huggingface.co/fse/glove-wiki-gigaword-300

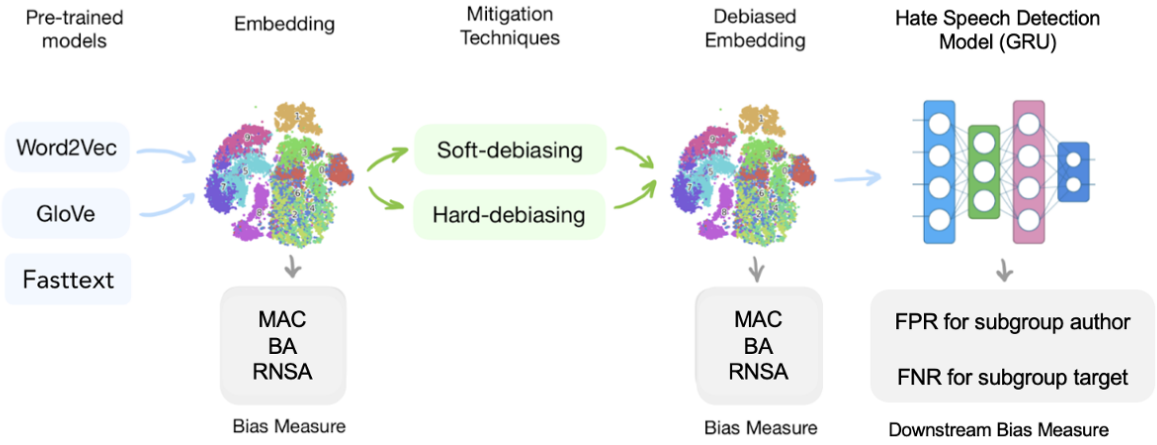

Figure 1: Methodological framework for hate speech detection, adapted from Fersini et al. (2023).

beddings, GloVe is tweet-based.

3.3 Model Selection and Training for Hate Speech Detection

Following Huang et al. (2020), who found that GRU performed reasonably well on their dataset, we use a Bidirectional GRU with pre-trained embeddings. Our model includes a dropout rate of 0.2, a sigmoid activation function, the Adam optimizer with a learning rate of 0.001, and is trained for 10 epochs using the cross-entropy loss function. For the modification of the embedding space, the embedding layer is frozen to avoid any dataset biases. Finally, the datasets are randomly split into training, validation, and test sets with ratios of 70%, 15%, and 15%, respectively.

3.4 Debiasing Embeddings

For our experiment, we evaluate two common debiasing methods for static pre-trained embeddings: Hard Debiasing and Soft Debiasing. Originally introduced by Bolukbasi et al. (2016) and later extended by Manzini et al. (2019)[6] for multi-class debiasing, both methods begin by identifying a bias subspace using Principal Component Analysis (PCA) on defined sets of identity terms (e.g., *he–she*). We use k=3 principal components for all embeddings except GloVe soft debiasing (k=1)[7].

Hard Debiasing (Neutralize and Equalize): This method begins by identifying the bias subspace using PCA on pre-defined sets of identity terms. For identity-neutral words (e.g., *doctor*, *nurse*), the component along this bias subspace is completely removed, ensuring they are equidistant from identity terms. For identity terms (e.g., *he*, *she* for gender), their embeddings are adjusted to be symmetrically positioned relative to neutral words, enforcing equal representation in the embedding space.

Soft Debiasing (Equalize and Soften, λ=0.2): Similar to hard debiasing, the bias subspace is identified using PCA. However, instead of fully removing the bias component, a linear transformation reduces its projection for gender-neutral words. The debiasing strength is controlled by λ, which balances the bias reduction and semantic preservation. A higher λ emphasizes stronger debiasing at the risk of distorting word relationships, while a lower λ retains the original structure more but reduces bias less effectively. Following Manzini et al. (2019), we select λ=0.2.

3.5 Bias Evaluation

We first evaluate the bias in the word embeddings themselves, to gauge the effect of the debiasing methods. We then evaluate the bias introduced into the hate speech detection model, pre- and post-debiasing.

3.5.1 Measuring Bias in Embeddings

We use three bias metrics in the embedding space, which capture different aspects: (1) stereotype bias in target group roles and (2) sentiment bias toward target groups. Stereotype bias is measured using Mean Average Cosine Similarity (MAC) and Bias Analogy (BA) while sentiment bias is as-

[6] https://github.com/TManzini/DebiasMulticlassWordEmbedding/tree/master

[7] For this setting, k=3 provides substandard results.

sessed using Relative Negative Sentiment Association (RNSA). See Table 7 in the appendix for the target and role terms used in BA and MAC calculations.

MAC (Manzini et al., 2019) measures differences among groups in relation to neutral terms. A MAC score[8] of 1 indicates that the term has no strong association with any identity group. A MAC score of less than 1 indicates presence of association bias with some identity group. For example, a MAC score of 1 between *male* and *female* for *nurse* suggests that *nurse* is not specifically associated with either gender. The MAC score is computed as follows:

$$\text{MAC} = \frac{1}{|T||A|} \sum_{T_i \in T} \sum_{A_j \in A} S(T_i, A_j)$$

where T and A represent target identity terms and neutral attribute terms, and $S(T_i, A_j)$ is the average of cosine distances between an identity term and a neutral term.

BA (Dev and Phillips, 2019) compares stereotypical associations of the target groups with their attributes, such as A is to [stereotype] as B is to [stereotype] (e.g., *male* is to *doctor* as *female* is to *nurse*). A higher BA score indicates stronger association; 1 implies strong stereotypical association, and 0 implies no association at all. BA is computed as follows:

$$\text{BA} = \frac{\sum_{T_i} \sum_{T_j} \sum_{A_n} \sum_{A_m} S(T_i, A_n, T_j, A_m)}{\binom{|T|}{2} |A||A'|}$$

where T are target identity terms, A are stereotypical attributes, and $S(T_i, A_n, T_j, A_m)$ is the cosine similarity between $T_i - A_n$ (e.g., *male* to *doctor*) and $T_j - A_m$ (e.g., *female* to *nurse*).

RNSA is our adaptation of Relative Negative Sentiment Bias (RNSB) (Sweeney and Najafian, 2019) to measure the contrast between an identity term's association with positive and negative sentiment words. RNSB calculates the KL divergence of sentiment distributions from a uniform distribution, with a value of 0 indicating no bias and higher values reflecting stronger sentiment association bias. RNSA, in contrast, focuses on the magnitude and the direction of sentiment bias (positive vs. negative), thus more directly indicating hate speech detection bias. Since RNSA is only applicable for binary groups, it is computed for gender bias only. A score of 0 indicates the target terms (i.e., *she*) are neutral or have no specific association with a certain sentiment, whereas 1 indicates positive sentiment and -1 negative. RNSA for any identity term is computed as follows:

$$\begin{aligned}\text{RNSA}(w) = &\frac{1}{|A_1|} \sum_{a_1 \in A_1} (1 - \text{cosine}(w, a_1)) \\ &- \frac{1}{|A_2|} \sum_{a_2 \in A_2} (1 - \text{cosine}(w, a_2))\end{aligned}$$

3.5.2 Bias in Hate Speech Detection

False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED), proposed by Dixon et al. (2018), are commonly used to assess bias in hate speech detection models (e.g., Park et al., 2018; Seshadri et al., 2022). These metrics measure False Positive (FP) and False Negative (FN) rates of texts containing different classes of identity terms. These texts often use synthetic text templates (Park et al., 2018), since real datasets are highly skewed, complicating direct bias evaluation. However, synthetic data, while controlling for confounding factors, fails to capture the bias in real-world settings.

To address this, our study employs datasets MTC-E (Huang et al., 2020) and SBIC (Sap et al., 2020) to measure the author bias and target bias with real posts:

Author bias computes FP by Author group to assess if minority groups are disproportionately flagged as hate speech authors.

Target bias computes FN by Target group to determine if hate speech against minority groups is overlooked.

While the target distribution in SBIC remains skewed, with hate disproportionately aimed at minority groups, the author distribution in MTC-E is balanced. Additionally, regression tests on the group coefficients remain asymptotically valid; the size of the significance tests[9] is still correct, while imbalance only reduces power.

[8] As a multi-class bias metric, MAC improves upon binary-class metrics like Word Embedding Association Test (Schröder et al., 2021), making it more suitable for analyzing multi-class race bias in our study.

[9] The Wald z-test is used for significance testing in logistic regression and generalized linear models (GLMs).

4 Results

4.1 How Does Debiasing Reshape Embeddings?

Table 1 shows the results of soft and hard debiasing on the different embeddings. The improved MAC scores on GloVe, BA scores on FastText, and RNSA on Skipgrams and Glove show that both hard and soft debiasing can effectively reduce stereotypical associations and sentiment bias. Soft debiasing proves to be more robust with respect to MAC, showing consistently reduced associations between identity terms and neutral words. Meanwhile, BA favors hard debiasing, indicating that hard debiasing consistently reduces stereotypical analogies. Sentiment bias, as measured by RNSA, does not seem to differentiate between the two debiasing methods.

As shown in Table 1, different identity categories exhibit distinct bias patterns, with gender showing less analogy bias (lower BA) but stronger identity-neutral word associations (lower MAC) than race. Correspondingly, debiasing is more effective for gender in reducing identity-neutral word associations (higher MAC) and for race in mitigating stereotypical analogies (lower BA). For instance, for soft-debiased Skipgrams, MAC increases by 0.083, indicating reduced identity-neutral word associations, but BA increases by 0.065, reinforcing stereotypical analogies. Conversely, for race, debiasing consistently improves BA scores, but slightly increases identity-neutral word associations, as seen in a minor MAC decrease of 0.001 in hard-debiased Word2Vec and FastText.

Notably, models exhibiting stronger stereotypical associations can encode low sentiment bias, indicating that stereotype and sentiment bias are not correlated. FastText, which gives the worst MAC and BA scores, exhibits the least sentiment bias (best RNSA), whereas GloVe, which appears less biased according to MAC and BA, embeds the highest sentiment bias (worst RNSA).

Aligning with Fersini et al. (2023), our results show that debiasing methods do not uniformly reduce bias across embeddings and bias types: (1) Hard debiasing tends to mitigate stereotypes (better BA) but reinforces identity-to-neutral word associations (worse MAC), especially for race; (2) Soft debiasing reduces identity-to-neutral word associations (better MAC) but reinforces stereotype analogies (worse BA); (3) Stereotype and sentiment bias are not correlated: FastText, despite stronger stereotypes, has lower sentiment bias, while GloVe, appearing fair under MAC and BA, embeds greater sentiment bias. These results underscore the need for bias-specific debiasing strategies.

4.2 How Does Debiasing Affect Hate Speech Detection?

To examine the impacts of debiasing methods on hate speech detection, we assess whether these methods: (1) reduce unjustified censorship against minority group authors (Author bias), and (2) improve the model's ability to detect hate directed at minority group targets (Target bias).

4.2.1 Author Bias Evaluation

Using MTC-E (Huang et al., 2020), we investigate whether models before and after debiasing exhibit author bias for gender (female vs. male) and race (non-white vs. white). We use a logistic model, with male and white being the reference group. The results are shown in Table 2.

Most pre-debiased models show no author bias against female and non-white (i.e., FP lower than the reference group). The GloVe model is the only setting with author bias, mislabeling more non-hate posts by female users as hate posts than those by male users, but the difference is not statistically significant. Meanwhile, all pre-debiased models are more likely to correctly detect hate speech posted by female and non-white authors than the reference group (i.e., lower FN than the reference group), especially for non-white ($p < 0.001$).

Regarding FP, we observe contrasting effects between the two debiasing methods: Hard debiasing reduced the FP rate for female authors over male authors, whereas soft debiasing aggravates both gender and race biases by increasing FP for female and non-white authors. One exception is FastText, as it decreases FP for non-whites and creates a marginally significant increase in FP for females ($p = 0.07$).

Both debiasing methods affect FN differently for the different biases. For gender bias, most post-debiasing models (all but soft-debiased GloVe) show a larger decrease in FN for female authors than male authors (e.g., hard-debiased Skipgrams coefficient -0.224, soft-debiased -0.248; $-.0251$ before debiasing), though not statistically conclusive. In contrast, both debiasing methods somewhat increase FN for non-white over white authors (except Skipgrams).

Embeddings	Debias	MAC ↑ Gen	MAC ↑ Race	BA ↓ Gen	BA ↓ Race	RNSA ↓ Gen
Skipgrams	orig	.813	.946	.147	.493	-.016
	hard	.823	*.945*	.139	.440	-.015
	soft	**.896**	**.964**	*.212*	*.509*	**.010**
FastText	orig	.592	.725	.532	.557	**.010**
	hard	.767	*.724*	.220	.456	*.011*
	soft	.695	.837	.220	.365	*.012*
GloVe	orig	.803	.911	**.095**	.402	.038
	hard	.840	.913	*.099*	.291	.028
	soft	.817	.958	*.122*	**.218**	.033

Table 1: Embedding bias evaluation across gender and race metrics. Bold: best scores; italics: worse after debiasing; ↑: 1 is fair; ↓: 0 is fair.

Emb		FP Gender	FP Race	FN Gender	FN Race
Skip	O	-.0204	-.1355	-.0251	*-.1973*
	H	.1280	.0938	-.0224	-.0278
	S	.0083	.0666	-.0248	-.0512
Fast	O	-.0747	-.0265	-.0347	**-.2790**
	H	.0513	-.1145	-.0231	.0335
	S	.2482	-.0048	-.0329	.0451
GLV	O	.0456	-.0945	-.0351	*-.2244*
	H	-.0001	-.0318	-.0196	.0525
	S	.0493	.0038	.0438	.0352

Table 2: Model bias evaluation on author bias using logistic model FP/FN ∼ identity × debias. pos/neg. numbers: higher/lower than the reference group. Significance: bold: ($p < 0.01$), italics ($p < 0.05$).

Emb		Gender Female	Race Black	Race Asian
Skip	O	-.4227	-.8449	-.0896
	H	*.6845*	-.3939	-.2299
	S	*.7023*	-.2906	-.4087
Fast	O	.2561	-.8262	.4249
	H	.4190	.9200	.7178
	S	-.1048	-.7079	-.8303
GLV	O	.5398	-.4327	-.7918
	H	-.0940	-.5220	.4722
	S	-.0361	-1.2010	-.5911

Table 3: Model bias evaluation on target bias using logistic model FN ∼ identity × debias. pos./neg. numbers: higher/lower than the reference group. Significance: italics ($p < 0.05$).

4.2.2 Target Bias Evaluation

We evaluate models, pre- and post-debiasing, using SBIC (Sap et al., 2020) to examine target bias for gender (male, female) and race (White, Black, Asian) in detecting implicit hate speech, setting male and White as the reference groups. Table 3 shows the results.

Before debiasing, all models except Skipgrams are more likely to miss hate speech targeting females compared to males (higher FN for female). For race, hate speech targeting Black and Asian individuals is less likely to be missed than hate speech against white individuals (lower FN). However, if Black is used as the reference group, hate speech targeting Asians becomes significantly more likely to be missed by the models ($p < 0.001$).

The two debiasing methods have different effects: Hard debiasing generally increases FN for the target group (female, Black, Asian), exacerbating bias (except for Skipgrams in race and GloVe in gender). This aligns with findings by Şahinuç et al. (2023) that hard-debiased FastText increases gender-related bias in hate speech detection. In contrast, soft debiasing, with the exception of Skipgrams, tends to reduce FN for the target group. We must note that none of these effects are statistically significant, with the exception of Skipgrams, which initially shows lower FN for hate against females (favoring female individuals), then a significant increase in FN after both hard- and soft-debiasings ($p = 0.041$ for hard debiasing; $p = 0.042$ for soft debiasing). Skipgrams, post-debiasing, become more likely to miss hate speech against females than males, though this difference remains insignificant. This observation confirms findings by Park et al. (2018) that debiased Word2Vec without data augmentation increases gender bias.

In summary, author and target bias respond dif-

Dataset		MAC	BA	RNSA
MTC-E	FPR	**-0.67**	0.02	**-3.55**
	FNR	**0.43**	0.01	**1.69**
SBIC	FPR	**-0.57**	**-0.28**	**0.82**
	FNR	**0.65**	**0.04**	**2.10**

Table 4: GLM: FPR/FNR $\sim$ metric. $+/-$ indicate increase/decrease with fairer metric values respectively. Significance: bold ($p < 0.001$).

Type		MAC	BA	RNSA
Gend.	Female			-5.25
Race	Asian	**-4.07**	**-3.23**	
	Black	*-1.84*	*-1.36*	

Table 5: GLM: FNR $\sim$ group $\times$ metric. $+/-$ indicate increase/decrease of FNR with fairer metric values. Significance: bold ($p < 0.001$), underline ($p < 0.01$), italics ($p < 0.05$).

ferently to debiasing methods, and their effects vary across identity groups. Hard debiasing tends to be more effective in reducing author bias, particularly for race. In contrast, soft debiasing is more effective in mitigating target bias.

5 Discussion

The inconsistency in debiasing effectiveness across different conditions in the embedding space analysis (Section 4.1) aligns with our findings wrt. model bias (Section 4.2). For example, soft debiased Skip-grams reinforce more gender than racial stereotypes and correspondingly increase the likelihood of missing hate speech against gender but not race. Given this, we examine the embedding space to investigate whether it provides a more coherent explanation for these distinctions (Section 5.1). We also perform an error analysis to illustrate the effect of embedding bias on hate speech detection (Section 5.2).

5.1 Embedding Space Bias and Model Bias

To analyze author bias and target bias, we apply generalized linear models (GLM) to models with both MTC-E and SBIC to examine the correlation between changes in embedding metrics and the general model performance across the two datasets. No significant association is found for MTC-E while significance is reported for SBIC. This suggests that embedding space bias metrics are reliable predictors for target bias but not author bias:

As shown in Table 4, for both datasets, improvements in MAC and RNSA are significantly correlated with a lower general false positive rate (FPR) but a higher false negative rate (FNR). This suggests that the models become less likely to label posts as hate. Although undesirable, this outcome is expected, as debiasing removes contextual information from the embedding space in exchange for bias removal. However, BA shows distinctive correlations between the two datasets: improvements in BA are significantly associated with improvements in both FPR and FNR for MTC-E, but with improved FPR and worse FNR for SBIC. This aligns with the finding that debiasing in general improved model performance for MTC-E, but not for SBIC (apart from FastText), which contains implicit hate and thus relies more heavily on contextual information.

While SBIC shows increased FNR as indicated by the metrics[10], the decrease in FPR indicates that the model bias reduced significantly. This demonstrates both a trade-off between model performance and model bias, and the necessity of combining embedding debiasing with other methods to both reduce bias and improve model performance, agreeing with Park et al. (2018).

Table 5 presents correlations between the change in FNR and the metrics by target group. As shown, no target group experiences a significant increase in bias following improvements in any metric. More importantly, all three metrics are positively associated with decreased FNR for the selected minority groups compared to the reference group, though the associations differ by identity category:

MAC & BA evaluate reduction of *stereotype bias* in the embeddings. Improvements in both metrics significantly correlate with reduced FNR for hate against Asian and Black targets compared to White, with a stronger effect for Asians. However, no significant association with gender bias was observed.

RNSA targets *sentiment associations* in embeddings for gender. The decrease in RNSA is significantly associated with reduced FNR for female targets compared to male targets.

Our results suggest that metrics quantifying bias in the embedding space can significantly predict the carried-over bias in hate speech detection. However, the choice of metric should align with specific

[10]Note that the model's general performance (F1) did not change significantly across models (general performance can be found in Table 8 in the appendix).

Bias type	Posts labeled as Offensive
Stereotype	*How do you know your house was robbed by an Asian? Your homework is done for you, your computer hardware has been upgraded, and the guy has still been trying to back out of your driveway for the past 30 minutes.* (biased GloVe: 0.4407, soft-debiased: 0.9987) *happy chinese new year enjoy the traffic free drive to work! (meant to post this yesterday)* (biased GloVe:0.0963, soft-debiased: 0.6199) *Why did Santa spend so much time in China last night He was picking up all the toys* (biased GloVe:0.4444, soft-debiased: 0.9637)
Sentiment	*making sweet love to my hand cause i cant buy a h*e tonight.* (biased Word2Vec: 0.1119, soft-debiased: 0.7800) *Well that's ok-you're a girl.* (biased Word2Vec: 0.1560, soft-debiased: 0.6936) *an older female english teacher asks. <repeat>. <repeat> "if i say 'i am pretty', i am speaking in which tense?" little johnny raises his hand and says, "obviously in the past".* (biased Word2Vec: 0.4460, soft-debiased: 0.9645)

Table 6: Top tweets from SBIC that moved to hate by debiased models.

bias reduction objectives, as different groups (e.g., race, gender) exhibit distinct patterns of model bias.

5.2 Error Analysis

Given that embedding metrics can reliably predict model bias, we have conducted an error analysis to assess how improvements in embedding space bias translate into hate speech detection outcomes. We compared the performance of FastText before and after soft debiasing for race (Asian), as it showed the greatest improvement in MAC and BA. Similarly, we analyzed the Skipgram model before and after soft debiasing for hate against female, which demonstrated the largest improvement in RNSA. We focused on posts that witnessed the greatest shift from non-hate to hate predictions after debiasing. The posts are shown in Table 6.

Our results indicate that, for race (Asian), debiasing leads to the detection of posts reinforcing harmful stereotypes, such as those associating Asians with being hard-working or bad drivers (e.g.,the first example). For gender, debiasing revealed posts with harmful associations, such as comments linking sentiment to women's physical attractiveness.

6 Conclusion and Future Work

This study explored the role of embedding bias and its impact on hate speech detection bias. Our findings highlight several critical conclusions:

Bias as Author vs. Target: Bias against minority groups differs when they are the *authors* versus the *targets* of hate speech. Embedding debiasing methods cannot effectively reduce author bias but show limited success for minority target groups.

Distinct Nature of Gender and Race Bias: The distinction between stereotype and sentiment metrics in predicting model bias for identity categories highlights that gender and race bias have fundamentally different characteristics, meaning approaches effective for one cannot be directly applied to the other: (1) MAC and BA effectively identify racial bias reductions for minority target groups, particularly Asians. (2) RNSA performs better for detecting gender bias improvements.

In conclusion, this work underscores the complexity of bias in hate speech detection. Bias varies across gender, race, and specific racial groups, necessitating group-specific approaches and nuanced bias metrics. A single debiasing method or metric cannot universally address all bias issues, promoting the need for targeted solutions and refined evaluation frameworks to achieve fairer hate speech detection systems.

We will continue our efforts to explore other model architectures, including SVMs (for explainability) and Transformer-based models. Additionally, we will investigate debiasing methods for contextualized word embeddings (e.g., Kaneko and Bollegala (2021); Zhao et al. (2019)), which can capture richer semantic and syntactic nuances, to determine whether we see the same interactions between debiasing and evaluation metrics.

7 Limitations

The major limitation of this work results directly from the datasets. Our work focused on English, and we only used one dataset with author information and one with target information. Ideally, this investigation should be carried out using a larger set of datasets and other languages.

References

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*, pages 4356–4364.

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

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR.

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*, pages 67–73.

Elisabetta Fersini, Antonio Candelieri, and Lorenzo Pastore. 2023. On the generalization of projection-based gender debiasing in word embedding. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 336–343, Varna, Bulgaria.

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 Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*.

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*, pages 615–621, Minneapolis, MN.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Dante Razo and Sandra Kübler. 2020. Investigating sampling bias in abusive language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 70–78, Online.

Furkan Şahinuç, Eyup Halit Yilmaz, Cagri Toraman, and Aykut Koç. 2023. The effect of gender bias on hate speech detection. *Signal, Image and Video Processing*, 17(4):1591–1597.

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*, pages 5477–5490, Online.

Sarah Schröder, Alexander Schulz, Philip Kenneweg, Robert Feldhans, Fabian Hinder, and Barbara Hammer. 2021. Evaluating metrics for bias in word embeddings. *arXiv preprint arXiv:2111.07864*.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337*.

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.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 602–608, Minneapolis, MN.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

8 Appendix

Gender

Target	Role (BA)	Neutral term (MAC)
he, his, son, father, male, boy, uncle	manager, executive, doctor, lawyer, programmer, scientist, soldier, supervisor, rancher, janitor, firefighter, officer	chair, house, supervisor, secretary, loud, weak
she, hers, daughter, mother, female, girl, aunt	secretary, nurse, clerk, artist, homemaker, dancer, singer, librarian, maid, hairdresser, stylist, receptionist, counselor	

Race

Target	Role (BA)	Neutral term (MAC)
caucasian, America	manager, executive, redneck, hillbilly, leader, farmer	chair, house, smart, criminal, executive, farmer
black, Africa	slave, musician, runner, criminal, homeless	
asian, Asia	doctor, engineer, laborer, teacher	

Table 7: Target terms with their stereotypical roles for BA, and neutral terms for MAC, following Manzini et al. (2019).

				Gender		Race	
Dataset	Emb	Class	Biased	Hard	Soft	Hard	Soft
MTC-E	Skip	Hate	0.7381	0.7328	0.7369	**0.7425**	0.7320
		Non-Hate	0.8487	0.8590	0.8580	**0.8613**	0.8511
	Fast	Hate	0.7380	0.7425	0.7320	**0.7465**	0.7358
		Non-Hate	0.8476	0.8440	**0.8545**	0.8478	0.8504
	GLV	Hate	**0.7386**	0.7347	0.7362	0.7366	0.7378
		Non-Hate	0.8424	0.8521	0.8498	0.8512	**0.8535**
SBIC	Skip	Hate	**0.8803**	0.8638	0.8644	0.8690	0.8688
		Non-Hate	**0.8572**	0.8455	0.8407	0.8479	0.8405
	Fast	Hate	0.8736	0.8768	**0.8781**	0.8727	0.8740
		Non-Hate	0.8408	**0.8516**	0.8455	0.8478	0.8494
	GLV	Hate	**0.8870**	0.8735	0.8706	0.8640	0.8704
		Non-Hate	**0.8614**	0.8474	0.8462	0.8454	0.8405

Table 8: Debiased embedding results comparing MTC-E and SBIC, reporting separate F1 scores for Positive and Negative classes per dataset, embedding, and debiasing method.