Evolving Fairness - Adaptive Reinforcement Learning for Gender Bias Mitigation in Word Embeddings

Mr. Taahaa Mir
McGill Univeristy
taahaa.mir@mail.mcgill.ca

Mr. Mohamed Mahmoud
McGill Univeristy
mohamed.mahmoud@mail.mcgill.ca

Abstract

Our project introduces a new and innovative framework for mitigating gender bias in word embeddings, namely through reinforcement learning. Word embeddings, crucial for tasks like sentiment analysis, machine translation, job recommendation systems, often inherit biases from training data, perpetuating stereotypes and affecting decision-making in AI applications. Addressing these biases is essential for ethical AI practices. Our framework, developed in a custom OpenAI Gym environment, utilizes adaptive RL strategies to dynamically interact with and adjust embeddings, aiming to reduce gender bias while maintaining semantic integrity.

The novel aspect of this project lies in its adaptive approach, which applies a variety of actions—soft debiasing, counterfactual data augmentation (CDA), and strategic inaction—depending on the state of the embeddings. This is the final approach we reached and this method allows the RL agent to learn from the environment and modify its strategies to optimize bias reduction and semantic preservation.

Preliminary findings indicate that our model effectively diminishes gender bias while maintaining the functional utility of embeddings (not always in all cases). The project showcases a promising direction in creating fairer AI systems by ensuring that word representations are both accurate and unbiased. Future work will focus on refining the reward function and expanding the approach to other forms of bias, enhancing the generalizability and effectiveness of bias mitigation in AI.

Introduction

Word embeddings are a fundamental component of Natural Language Processing (NLP). They convert text into numerical vectors that machines can process, capturing intricate patterns of language syntax and meaning within a compact vector space. These vectors are crucial across numerous NLP applications, including sentiment analysis and machine translation. However, the development of these models is not free from challenges. One significant concern is the inadvertent encoding of societal biases—especially gender biases—during the training phase. When not addressed, these biases may carry over into real-world applications, potentially influencing decision-making systems in biased ways. This issue is highlighted by our analysis of popular pretrained embeddings. We projected these embeddings onto a gender direction and visualized their position with respect to the gender direction:

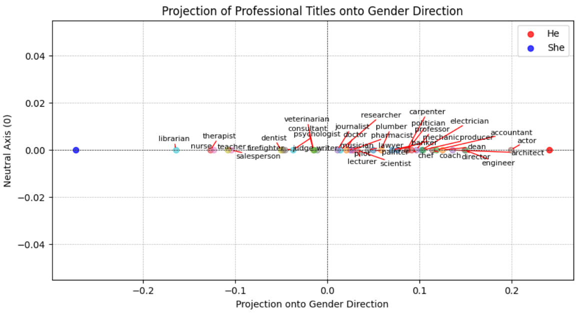


Figure 1: Glove Embedding Bias Analysis

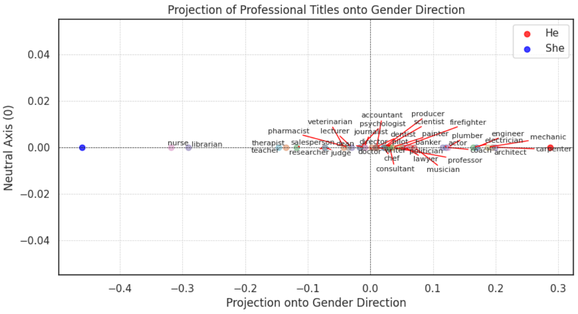


Figure 2: Word2Vec Bias Analysis

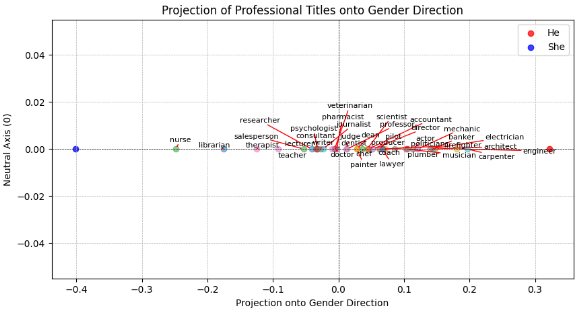


Figure 3: FastText Bias Analysis

It is clear that each of the embedding showed signs of bias, neutral professional roles are clearly more aligned with a certain gender in all the graphs above. These data sets were obtained from: (Pennington et al., 2014), (Grave et al., 2018), (Norwegian Language Processing Laboratory, 2024)

We further analyzed the biases by generating a cosine similarity heatmap with each embedding and the gender direction Figure (4, 5, 6):

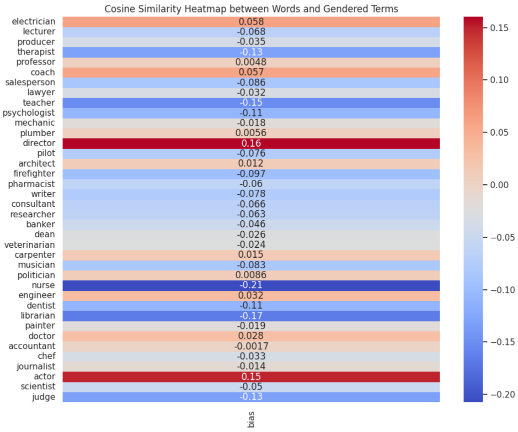


Figure 4: Cosine Similarity: Glove Embedding with Gender Direction

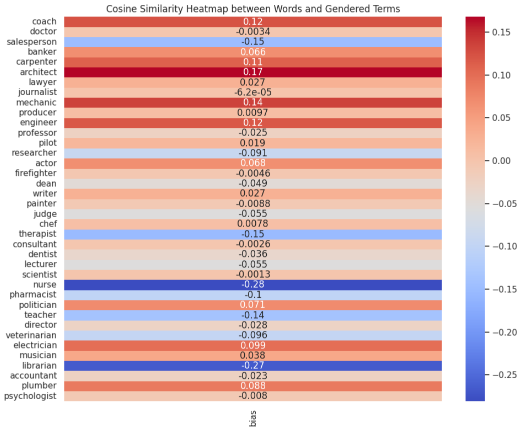


Figure 5: Cosine Similarity: Word2Vec Embedding with Gender Direction

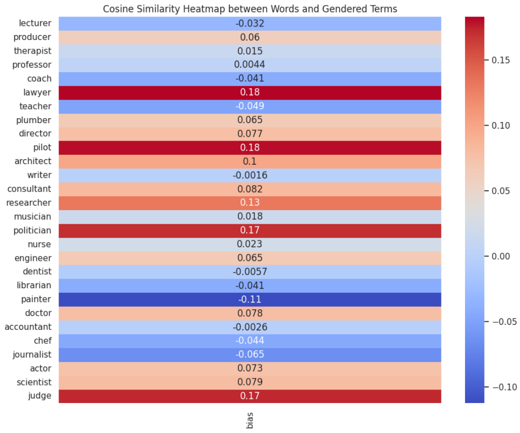
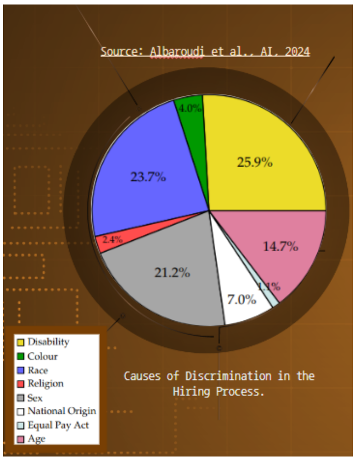


Figure 6: Cosine Similarity: FastText Embedding with Gender Direction

The consequences of biases in word embeddings extend beyond theory and have tangible effects on society. Take, for example, the study by a legal firm shown in Figure 7, which highlights the widespread issue of bias in recruitment processes. This study underscores how unchecked biases in AI can exacerbate social inequalities. These concerns are especially acute in vital sectors like employment, where biased decision-making can affect diverse demographics, cutting across lines of gender, ethnicity, and age. Our research introduces an adaptive reinforcement learning (RL) framework that is designed to evolve as language does. Moving away from static models, this dynamic framework integrates the fluid nature of language, enabling ongoing adjustments to word embeddings that reflect a contemporary understanding of fairness. By leveraging modified Q-Learning and Deep Q-Networks (DQN), our



Source: Albaroudi et al., AI, 2024

Figure 7: Causes of Discrimination in the Hiring Process

model applies a trio of targeted actions, each contributing to the nuanced task of reducing bias.

By dynamically adjusting word embeddings, our system aims to mitigate gender bias observed in cosine similarity heatmaps and projections of professional titles onto the gender direction. Our method carefully balances the trade-off between reducing bias and preserving the semantic richness of the embeddings, aligning with our overarching objective of cultivating AI that is not only linguistically adept but also ethically conscious. With this research, we aspire to lay the groundwork for AI technologies that advocate for inclusivity and uphold the values of a fair and equitable society.

Related Works

In the domain of mitigating bias within AI systems, especially in word embeddings, substantial progress has been made in recent years.

Bias in Pre-trained Embeddings

(Sesari et al., 2022) conducted an extensive empirical analysis on fifteen widely-used pre-trained word embedding models generated from algorithms like GloVe, word2vec, and fastText. Their research employed four distinct bias metrics: WEAT, SEMBIAS, DIRECT BIAS, and ECT. Findings from their study indicated that fastText models were the least biased in a majority of cases.

Notably, their analysis also revealed that models with smaller vector sizes tended to exhibit increased bias. This insight is crucial for our approach as it underscores the importance of considering embedding dimensions when developing bias mitigation strategies, influencing our decision to integrate adaptive mechanisms that can respond to varying embedding characteristics dynamically. We will assess our algorithm on vectors with relatively lower sizes (100 dimensions) to assess the mitigation of bias in this critical scenario.

Additionally, the work by (Caliskan et al., 2017) on measuring social biases in sentence encoders is particularly relevant. Their introduction of the Word Embedding Association Test (WEAT) provides a foundational methodology for detecting and quantifying biases embedded in text data. WEAT measures semantic similarity between sets of target words and attribute words, allowing researchers to assess bias based on how embeddings associate these concepts. This approach aligns with our project's objective of mitigating gender bias in word embeddings and is therefore used by us in our bias analysis.

Advanced Mitigation Techniques

Mitigation of bias in word embeddings has traditionally involved methods such as those introduced by (Bolukbasi et al., 2016), which aimed to adjust embeddings to reduce gender stereotypes while preserving the desired semantic associations, a version of their approach is utilized in our algorithm. Extending beyond these foundational techniques, our research integrates dynamic elements like Counterfactual Data Augmentation (CDA). As explored by (Zmigrod et al., 2020), CDA employs linguistic transformations to balance gender representations in the training data, thereby reducing learned biases.

Building upon these foundational techniques, our research delves into more complex methods for bias mitigation. Our approach leverages the foundational methodologies and introduces an adaptive reinforcement learning framework that continuously refines and optimizes the debiasing process, accommodating real-time data feedback and evolving linguistic patterns.

Reinforcement Learning for Bias Mitigation

The innovative use of reinforcement learning (RL) for bias mitigation marks a promising development in this field. Drawing inspiration from the work of (Ashioya, 2024), who advocated for RL algorithms like RLHF and RLAIF to address gender bias in AI systems, our project implements the suggestion with RLAIF allowing for ongoing adaptations to the embeddings with bias and semantic feedback at each step. This approach is particularly important given the critique by (Gonen and Goldberg, 2019) that biases are deeply embedded within language models and cannot be fully addressed through static methods alone. Our RL framework is designed to dynamically interact with the embedding space, making incremental adjustments based on continuous learning and feedback, thus ensuring that our mitigation efforts are both robust and responsive.

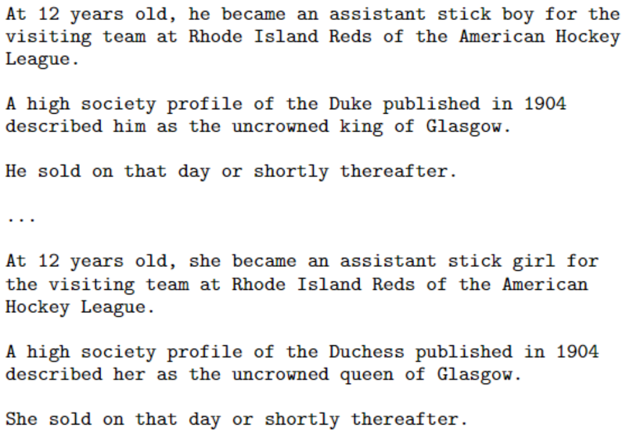
Broader Contextual and Causal Approaches

(Yang and Feng, 2019) study introduced the use of causal inference methods to deepen our understanding of how gender-defined word embeddings relate to gender-biased embeddings. This transition from associative analysis to a causal framework allows for more precise and targeted debiasing interventions. Such an approach complements our reinforcement learning model's ability to identify and modify both essential and unnecessary gender associations within embeddings. This advancement is pivotal as it enhances our toolkit for addressing biases, enabling a more detailed examination that takes into account not only the linguistic context but also broader societal trends, as demonstrated by (Garg et al., 2018) in their study.

By integrating these approaches, our project addresses both the technical aspects of bias mitigation and the broader societal and ethical implications of deploying AI technologies. Our use of adaptive reinforcement learning, combined with established debiasing techniques, aims to develop AI systems that are not just fair and equitable, but also responsive to shifts in societal values and linguistic norms. This holistic approach ensures that our AI systems can adapt and remain relevant as social dynamics evolve.

Dataset and Evaluation Metrics

For examples of the datasets, here is the dataset for CDA, For Common Crawl see (Appendix 18)



```
At 12 years old, he became an assistant stick boy for the
visiting team at Rhode Island Reds of the American Hockey
League.

A high society profile of the Duke published in 1904
described him as the uncrowned king of Glasgow.

He sold on that day or shortly thereafter.

...

At 12 years old, she became an assistant stick girl for
the visiting team at Rhode Island Reds of the American
Hockey League.

A high society profile of the Duchess published in 1904
described her as the uncrowned queen of Glasgow.

She sold on that day or shortly thereafter.
```

Figure 8: Counterfactual Data Augmentation Dataset

Data Acquisition

The primary dataset used in this study is sourced from the Common Crawl corpus via the Hugging Face dataset hub, (KeiRP, 2024) This dataset provides a broad representation of the internet's text, making it suitable for training robust word embeddings that capture a wide variety of linguistic contexts and usages.

Counterfactual Data Augmentation (CDA)

Additionally, we incorporate a specialized dataset for Counterfactual Data Augmentation (CDA) derived from (Currey et al., 2022). This dataset offers counterfactual examples where gender indicators in sentences are alternated, providing a balanced representation of gender in various contexts. This data is crucial for training embeddings that are less biased from the outset, addressing gender bias proactively.

Embedding Training

Both datasets are utilized to train FastText (Bojanowski et al., 2017) models, which are known for their efficiency in capturing subword information and handling out-of-vocabulary words. The process involves:

1. **Preprocessing**: Text data from the Common Crawl sample and the MT-GenEval dataset are cleaned and prepared. This includes normalizing text and ensuring it is suitable for training the embeddings.

2. **Training the FastText Model**: We train separate FastText models on:

   - Normal corpus derived from Common Crawl for general-purpose embeddings.

- Counterfactual data for creating embeddings that inherently possess reduced gender biases.

3. Each trained model outputs embeddings that are then normalized to ensure they are suitable for use in our RL environment.

Dataset Statistics

Here is a table of statistics that provides a detailed overview of the datasets used:

Dataset	Lines	Avg Length
Common Crawl	5,505	436.17 words
MT-GenEval (CDA)	8,400	22.64 words

Table 1: Statistics of datasets used in the study.

Gender	Number of Lines
Female	4,200
Male	4,200

Table 2: Number of lines where the subject is 'Male' vs 'Female' in the MT-GenEval (CDA) dataset.

Evaluation Metrics

We focus on two key aspects for evaluating the efficacy of our approach, ensuring both the reduction of bias in embeddings and the retention of their utility for NLP tasks:

1. **Bias Measurement**: Bias in word embeddings is quantitatively assessed by projecting word vectors onto a predefined gender direction vector, both before and after the application of debiasing strategies. The gender direction is typically constructed by calculating the difference between vectors for gender-specific words like "he" and "she". The projection quantifies how closely a word's representation aligns with gendered conceptual axes, enabling us to measure the effectiveness of our debiasing interventions quantitatively. This projection is mathematically defined as:

$$\text{Projection} = \frac{\vec{v} \cdot \vec{d}}{||\vec{d}||}$$

where $\vec{v}$ is the word vector and $\vec{d}$ is the gender direction vector. A reduction in this projection metric post-debiasing indicates successful bias mitigation.

2. **Semantic Integrity**: To ensure that the debiasing process does not compromise the embeddings' utility, we measure the cosine similarity between the original and adjusted embeddings. High cosine similarity values post-debiasing indicate that the linguistic utility of the embeddings is preserved. Cosine similarity is given by:

$$\text{Cosine Similarity}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| ||\vec{b}||}$$

Maintaining high cosine similarity is crucial as it ensures that while biases are reduced, the functional integrity of word embeddings for tasks like sentiment analysis, machine translation, or content recommendation remains intact.

3. **WEAT (Word Embedding Association Test)**: The WEAT test assesses bias by comparing the relative similarity of two sets of target words (e.g., male and female names) to attribute sets (e.g., career and family terms). A significant differential in association strengths between the two sets indicates the presence of bias. This test provides a statistical measure of bias in the embeddings and helps validate the efficacy of the debiasing techniques used. The WEAT score is derived from the difference in mean association strengths, standardized against the standard deviation of all association differences.

This rigorous approach to evaluation allows us to finely balance the need for reducing bias within the word embeddings with the necessity of preserving the semantic content of embeddings by making sure that the sentence stills make sense after modifications, aligning with our broader objective of fostering fairness without sacrificing performance in AI-driven applications.

Training FastText Models

The FastText models are trained as follows:

1. **Training Process**: The models are trained using the skip-gram architecture, optimizing for context prediction which enhances the quality and relevance of the embeddings for Natural Language Processing tasks.

2. **Normalization and Usage**: Post-training, embeddings are normalized to unit length to

standardize their usage in the Reinforcement Learning environment, ensuring consistent performance across various tasks.

Models

Initial Model: Iterative Embedding Adjustment via Modified Q-Learning RL

Our initial approach tackles gender bias in word embeddings by iteratively adjusting the vector representations. We utilize a modified version of the Q-Learning algorithm, a cornerstone of reinforcement learning (RL), known for its efficiency in learning policies for sequential decision-making problems.

State Space

In our RL framework, the state is a representation of the current condition of the embeddings. It encapsulates the complex structure of word vectors and quantifies bias using the Word Embedding Association Test (WEAT), a well-established measure in computational linguistics for detecting bias. The state is thus a high-dimensional vector reflecting the bias score associated with each word in the embedding space, providing a nuanced view of the embeddings' current fairness.

Actions

Our agent's actions are carefully crafted to address bias while maintaining the integrity of the original word meanings. It operates as follows:

- An adjustment vector, computed as per the bias direction and magnitude, is applied to the embeddings. This vector is designed to nudge the embeddings towards a less biased state without compromising their semantic relationships.

- To ensure the relevance and efficacy of these adjustments, they are scaled proportionally to the state's context. This scaling factor ensures that the adjustment is significant enough to effect change while preventing semantic drift.

Reward Function

The reward function, denoted as R, is a critical component that guides the agent towards optimal policy formulation by quantifying the effectiveness of each action taken.

It is defined mathematically for a state transition as follows:

$$R(current\_bias, next\_bias) = \begin{cases} current\_bias - next\_bias, & \text{if } next\_bias \leq current\_bias, \\ -2 \times (next\_bias - current\_bias), & \text{if } next\_bias > current\_bias. \end{cases} \quad (1)$$

where:

- $current\_bias$ is the bias measurement of the current state as determined by WEAT, offering a precise snapshot of the embeddings' bias at a given timestep,

- $next\_bias$ is the bias measurement of the subsequent state following an action, which reflects the impact of the adjustment.

The reward function is strategically devised to incentivize reductions in bias, with positive rewards for actions that lower the bias score. Conversely, it imposes a punitive response, in the form of negative rewards, for actions that inadvertently intensify bias, thereby aligning the agent's learning trajectory with the objective of bias minimization.

Final Model Architecture

As discussed in the paper (Mir and Mahmoud, 2024), the model operates within a custom simulation environment, which is part of the OpenAI Gym toolkit—a platform used for developing and comparing reinforcement learning algorithms. This environment, called `EmbeddingDebiasingEnv`, manages word embeddings and evaluates the impact of various debiasing actions.

Components of the Environment:

- **State Space**: This includes the current word's embedding vector, a measurement of bias, a measure of semantic similarity to the original embedding, and the last action taken. Each component is represented as a multi-dimensional vector that encapsulates different attributes of the word's representation.

  Let e_w be the embedding vector of the current word w, with a dimension of d. The state space $\mathcal{S}$ for each word in the environment can be represented as:

$$\mathcal{S} = \begin{bmatrix} e_w \\ \text{bias}_w \\ \text{sim}_{w,o} \\ a_{\text{prev}} \end{bmatrix}$$

where:

  - $e_w \in \mathbb{R}^d$ is the embedding vector of the current word w.
  - $\text{bias}_w \in \mathbb{R}^{100}$ is a vector filled with the current bias measurement for word w.
  - $\text{sim}_{w,o} \in \mathbb{R}^{100}$ is a vector representing the semantic similarity between the current embedding of w and its original embedding.
  - $a_{\text{prev}} \in \mathbb{R}^{100}$ is a vector indicating the last action taken in the environment, represented numerically.

  This state representation captures all necessary features for our model to make decisions about actions to modify the embedding of the word in order to reduce bias while maintaining semantic integrity.

- **Model Actions:** The agent in our environment can take one of the following three actions to adjust the embedding towards reducing bias while attempting to preserve semantic integrity:

1. **Soft Debiasing**: This action modifies the embedding to diminish gender bias. It involves slightly adjusting the direction of the word's vector in the embedding space to reduce its projection onto the gender direction. This method is effective for general debiasing tasks but carries a risk of over-correcting, which could neutralize necessary gender-specific terms.

2. **Counterfactual Data Augmentation (CDA)**: This action adjusts the embedding by aligning it closer to a counterfactually augmented version. CDA is performed by altering the training data to create versions of data points where gender-specific words are swapped, thus providing a model with balanced representations of gender. This helps in maintaining the nuances and meanings of the original text but might still inherit biases present in the training data.

3. **Do Nothing**: This action maintains the current state of the embedding without any changes. It is employed to test the hypothesis that some embeddings are already optimally debiased and do not require further adjustment. This allows the system to avoid unnecessary modifications that might degrade the utility or accuracy of the embeddings.

Rationale for Actions: Including these three distinct actions allows the model to leverage the strengths of both active debiasing techniques while mitigating their potential drawbacks:

- Soft debiasing is adept at reducing overt biases but may inadvertently alter essential gender-specific terms, which are sometimes necessary for accurate language representation.

- CDA preserves the original meaning and nuances of the text but may not completely eliminate biases if the training data itself is biased.

- The 'Do Nothing' action provides a baseline to assess whether adjustments are necessary, thereby minimizing unnecessary interventions in the embeddings.

Reward Function

The effectiveness of an action is quantified through a reward function, which balances two objectives:

- Reducing gender bias in the embedding.

- Preserving the original semantic meaning of the word.

$$R(s,a) = \begin{cases} \gamma & \text{if } a = \text{'Do Nothing'} \\ -\gamma & \text{if } a = \text{'Do Nothing'} \\ 0.6 \times (\text{Bias Reduction}) - 0.4 \times (\text{Semantic Change}) & \text{otherwise.} \end{cases} \quad (2)$$

Where:

- Bias Reduction is the reduction in gender bias, which is calculated as the difference in bias before and after the action, normalized over the range of bias changes observed during training.

- Semantic Change measures the change in semantic similarity to the original embedding, also normalized to account for the varying degrees of semantic shifts that can occur.

- γ is a small positive reward or penalty for the 'Do Nothing' action, encouraging the agent to maintain a near-optimal state without unnecessary adjustments.

The 'Do Nothing' action is crucial when the current state is already near the desired unbiased state, hence avoiding any further adjustments that could potentially disrupt the embedding's utility. Conversely, a penalty is imposed when no action is taken, but an adjustment is required, nudging the agent towards taking corrective measures.

Utilizing Deep Q-Networks (DQN) in Our Model

Deep Q-Networks (DQN) (Sutton and Barto, 2018) represent a significant advancement in reinforcement learning, combining traditional Q-learning with deep neural networks.

Basics of Deep Q-Network (DQN)

In Q-learning, one of the foundational techniques in RL, the agent learns a *Q-value* for each action in each state, which estimates the total reward that can be obtained from taking that action in that state, followed by the best possible future actions.

Deep Q-Networks enhance Q-learning by using a deep neural network to approximate the Q-value function. The key advantages of using DQN include:

- The ability to handle high-dimensional state spaces, making it suitable for complex problems like word embedding debiasing where the state involves high-dimensional vectors.

- Improved convergence properties over standard Q-learning, due to the generalization capabilities of neural networks.

Application of DQN in `EmbeddingDebiasingEnv`

In our model, the DQN is used to determine the optimal action to take for each state in the environment to reduce bias in word embeddings. The process is as follows:

1. The **state** is represented by the current word's embedding vector, its associated bias, the semantic similarity to the original embedding, and the last action taken.

2. The **action space** includes three actions: Soft Debiasing, Counterfactual Data Augmentation (CDA), and doing nothing.

3. The **reward function** quantifies the effectiveness of actions by balancing bias reduction against semantic similarity loss.

When the agent takes an action, the DQN predicts the Q-values for each possible future action from the new state, allowing the agent to choose the action that maximizes expected rewards. This prediction involves forward passing the current state through the DQN model to obtain Q-value estimates.

Training the DQN

The DQN is trained iteratively:

1. The agent observes the current state, selects an action based on the DQN's predictions, and receives a reward from the environment.

2. The resulting new state and the reward are used to update the DQN. The update is performed using *backpropagation*, where the neural network's weights are adjusted to minimize the prediction error in Q-values, improving the accuracy of the agent's decision-making process.

3. Over time, as the DQN experiences more states and outcomes, it becomes better at predicting the actions that will lead to bias reduction in embeddings.

The use of DQN in the `EmbeddingDebiasingEnv` marks a sophisticated approach to mitigating bias in NLP, pushing forward the boundaries of what is possible with AI in ensuring fairness and equity in language models.

Advantages Over Baseline Methods

Compared to traditional debiasing techniques, which often involve simple heuristic adjustments, our model adapts dynamically. It learns which actions lead to the best balance of reduced bias and preserved meaning based on feedback. This adaptability allows it to effectively handle a wide range of embedding types. In addition, this approach would reward actions where the embeddings are improved and penalized when the embeddings are worsened. That way our final debiasing sequences of actions aim to contain the advantages of both methods and leave out the disadvantages.

Visual Representation

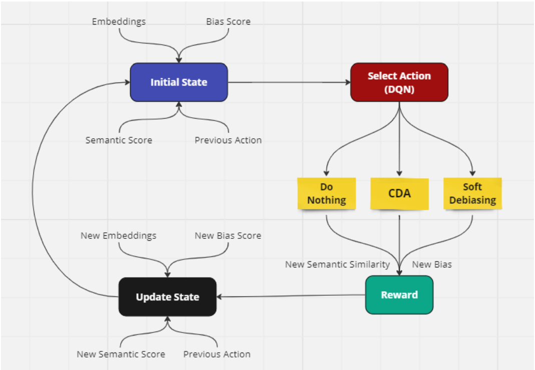


Figure 9: Model Architecture and Interaction

To aid understanding, we include a schematic diagram of the model's architecture and operational flow. Figure 9 illustrates the cyclical process where the agent selects actions, the environment updates the state and computes rewards, and the agent refines its strategy based on feedback. This iterative process continues until the embeddings are sufficiently debiased.

Baselines

The evaluation of our proposed debiasing approaches requires a robust comparison against established baselines. These baselines serve as a reference point to demonstrate the efficacy and advancements of our methods. We employ two primary baselines: traditional metrics and visual representations before and after debiasing interventions.

WEAT as a Benchmark

The Word Embedding Association Test (WEAT) serves as our primary quantitative baseline. This test measures the association of word embeddings with gendered terms before and after applying debiasing techniques. A significant shift towards neutrality in WEAT scores post-debiasing is indicative of successful bias mitigation.

The Word Embedding Association Test (WEAT) is a statistical measure that quantifies bias in word embeddings by testing whether there is a significant difference in associations between two sets of target words and two sets of attribute words. For example, given two sets of target words X and Y (e.g., words related to science and art, respectively), and two sets of attribute words A and B (e.g., words related to male and female gender), WEAT measures the association of X with A versus B and Y with A versus B. The hypothesis is that if X is more associated with A than B, and Y is more associated with B than A, then there might be a gender bias encoded in the embeddings.

The test statistic is defined as:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where $s(w, A, B)$ measures the association of a single word w with the sets of attribute words:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

with $\vec{w}$, $\vec{a}$, and $\vec{b}$ being the vector representations of word w, attribute word $a \in A$, and attribute word $b \in B$ respectively, and $\cos(\cdot, \cdot)$ is the cosine similarity between the vectors.

The effect size of this test is calculated to measure the separation between the distributions of the association scores for both sets of target words:

$$d = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(w, A, B)}$$

A large effect size indicates a larger difference in the association with the attribute sets, suggesting potential bias.

Visualization Baselines

We utilize two visualization techniques to qualitatively assess the biases in embeddings: gender projection plots and cosine similarity heatmaps.

Gender Projection Plots

These plots depict the positioning of professional titles along a gender direction, both before (Figure 10)
and after soft debiasing (Figure 11) and after CDA (Figure 12).
The baseline condition in Figure 11 shows signs of overdebiasing as even words such as 'he' and 'she' are gender neutral now. On the other hand, Figure 12, although the words have maintained the semantic meaning, there is still some bias.

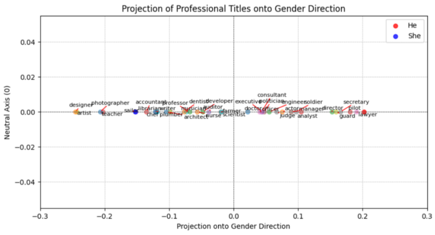


Figure 10: Embeddings before debiasing

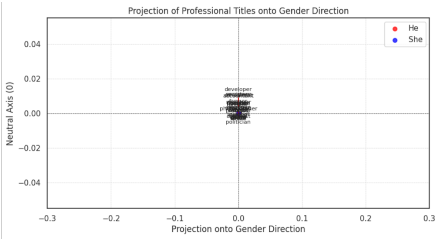


Figure 11: Embeddings after soft debiasing

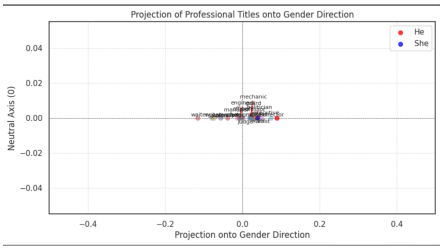


Figure 12: Embeddings after CDA

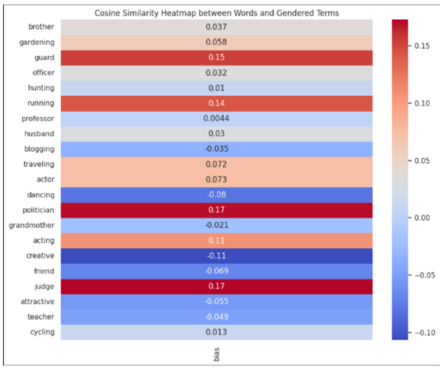


Figure 13: Cosine similarity of embeddings with gender direction before any debiasing

Cosine Similarity Heatmaps

Cosine similarity heatmaps provide a snapshot of the semantic closeness between words and gendered terms.

Our baseline heatmap (Figure 12) shows distinct biases, which we compared against the post-debiasing heatmaps to evaluate the preservation of semantic integrity alongside bias reduction.

Justification for Baselines

The chosen baselines are instrumental in demonstrating the initial state of bias within embeddings and the subsequent impact of our debiasing efforts.

They are particularly suited to validate our hypothesis that adaptive reinforcement learning can effectively mitigate bias while maintaining the utility of the embeddings for NLP tasks.

Furthermore, they provide a comprehensive evaluation framework by combining numerical and visual indicators of bias.

Experimental Details

In this section, we detail the experimental setup, including the hyperparameters used, the number of training epochs, the number of runs, and the software and libraries involved in the implementation of our model.

Hyperparameters

The model was configured with the following hyperparameters:

- Learning rate for DQN: 0.01
- Exploration fraction: 0.1
- Initial exploration probability: 1.0
- Final exploration probability: 0.1
- Number of training timesteps: 10000

These parameters were chosen to balance the tradeoff between exploration and exploitation during the training process.

Training Epochs

The training was structured into episodes. In reinforcement learning, an episode represents a sequence from the start to the termination of an environment. We trained the model for 100 episodes to ensure comprehensive learning across various states and actions.

Software and Libraries

The experiments were conducted using the following software and libraries:

- **Python**: The programming language used to implement the model.

- **OpenAI Gym**: Provided the custom environment `EmbeddingDebiasingEnv` for the RL agent.

- **Stable Baselines3**: An improvement of Stable Baselines, implemented in PyTorch, was utilized to run the DQN algorithm.

- **NumPy**: Employed for high-level mathematical functions.

- **Matplotlib and Seaborn**: Used for generating plots to visualize the training progress and results.

- **Scikit-learn**: The PCA functionality from this library was used to reduce the dimensionality of word embeddings for visualization purposes.

- **FastText**: Used to train word embeddings on the Common Crawl data.

- **AdjustText**: A library to make adjustments to text positioning in matplotlib plots to minimize overlaps.

The code for the experiments is available in our repository [https://github.com/MohamedAlKayd/Final-Project].

Results and Discussion

Baseline Comparisons

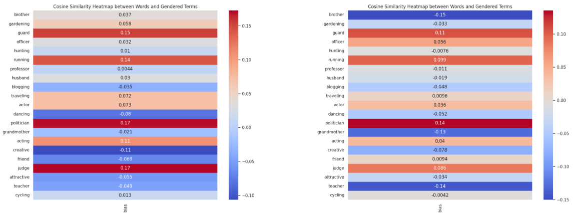
We established baselines using the pre-debiasing state of our embeddings and compared them with the results after applying our debiasing methods. The baseline embeddings showed significant issues with bias reduction as highlighted earlier.

Test **Pre-Debiasing (mean ± std dev)** **Post-Debiasing (mean ± std dev)**	
Pre-Debiasing	-0.4 ± 0.001
Post-Debiasing	-0.2 ± 0.001

Table 3: Performance of the baseline models and our proposed model.

Cosine Similarity and Gender Projections

Cosine similarity heatmaps before and after (Figure 14) debiasing illustrate the expected changes in associations between words and gendered terms. Our debiasing methods adjusted these associations towards neutrality, evident in the after-debiasing heatmap. Notably, terms like 'grandmother' retained their meaningful associations, while 'grandfather' demonstrated over-debiasing, shifting undesirably towards a female association.



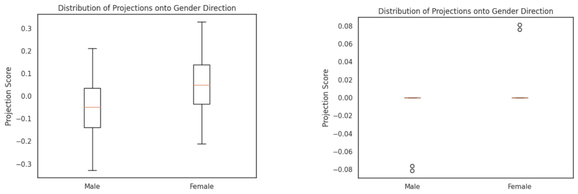
(a) Cosine Similarities with Gender Direction before RL

(b) Cosine Similarities with Gender Direction after RL

Figure 14: Cosine similarity and Gender Projections

Boxplot Analysis

The boxplot visualization (Figure 15) complements these findings by presenting the distribution of projections onto the gender direction. Post-debiasing, the medians converge, suggesting a shift towards gender neutrality. The Mann-Whitney U test, yielding a p-value of 0.0079, confirms the statistical significance of the observed changes. However, the presence of outliers underscores the complexity of debiasing and the need for refined approaches but still supports our results.



(a) Before our RL Debiasing

(b) After our RL Debiasing

Figure 15: Boxplot of distributions of projections onto the gender direction

Embedding Comparisons

Direct comparisons of word projections onto the gender direction before and after (Figure 16) debiasing reinforce these results. The after-debiasing projections indicate a shift towards the neutral axis, underscoring the effectiveness of our RL debiasing approach.

Reward Efficiency

Our reward function's efficiency is visualized in the reward graph (Figure 17). While the algorithm consistently worked towards reducing bias, as indicated by the overall reward trend, fluctuations

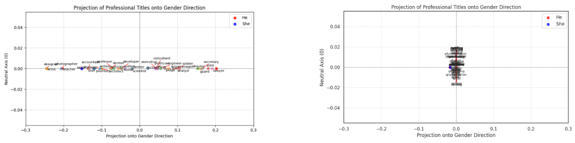


Figure 16: Embeddings before and after

suggest areas for potential improvement in reward function optimization.

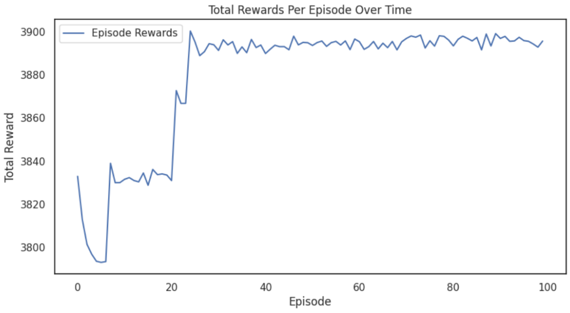


Figure 17: Reward Function Efficiency

Conclusion

In this research, we confronted the challenge of gender bias in word embeddings head-on with an innovative reinforcement learning approach. Our method works by dynamically adjusting the word embeddings within an environment that simulates the nuanced landscape of linguistic representation. One of the main aims was to preserve semantic integrity while methodically steering the embeddings away from inherent societal biases, towards a more gender-neutral space. The core of our strategy rested upon a bespoke reward function tailored to promote bias mitigation without compromising the fidelity of the embeddings' original linguistic intent. The results were promising. They demonstrated that we were able to successfully actively mitigate bias through an iterative, reinforcement learning-based process. Our model showcased an ability to distinguish between inherent gender associations necessary for language understanding and arbitrary biases that skew perception. This was evidenced by the balanced treatment of terms like 'grandmother' and 'grandfather' within the embeddings after debiasing.

As we look to the near future, several avenues for work present themselves. The exploration of other forms of bias, such as racial or age-related biases, within embeddings, is a natural progression of this study. Additionally, deploying our model in real-world applications will test its robustness and adaptability to the ever-evolving use of language. Moreover, as language models become more prevalent, our work underscores the imperative for ethical oversight in AI development, ensuring that these powerful tools serve to reflect and uphold the diverse and inclusive nature of human society.

In conclusion, our study takes a significant step towards responsible AI for NLP by presenting a method that not only detects but also actively corrects for gender bias in word embeddings. This endeavor is not merely an academic exercise but a commitment to the advancement of AI that is fair, accountable, and transparent, ultimately contributing to a more equitable digital future and better capabilities for Natural Language Processing.

References

Victor Ashioya. 2024. Using reinforcement learning algorithms to mitigate gender bias.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

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.

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

Anna Currey, Maria Nădejde, Raghavendra Pappagari, Mia Mayer, Stanislas LAULY, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *EMNLP 2022*.

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).

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

KeiRP. 2024. Common crawl sample. https://huggingface.co/datasets/keirp/common_crawl_sample.

Taahaa Mir and Mohamed Mahmoud. 2024. Adaptive reinforcement learning for mitigating gender bias in natural language processing. Submitted to McGill University in Partial Fulfillment of Requirements for Final Project for Graduate Course COMP579: Reinforcement Learning.

Norwegian Language Processing Laboratory. 2024. NLPL word embeddings repository. Available from NLPL at: http://vectors.nlpl.eu/repository/. Accessed: 2024-04-10.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. Available from Stanford NLP at: https://nlp.stanford.edu/projects/glove/. Accessed: 2024-04-10.

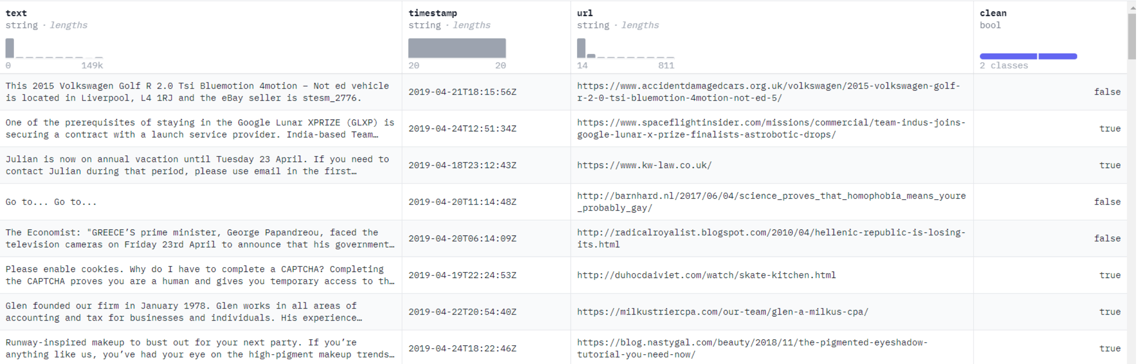
Emeralda Sesari, Max Hort, and Federica Sarro. 2022. An empirical study on the fairness of pre-trained word embeddings. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 129–144, Seattle, Washington. Association for Computational Linguistics.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Zekun Yang and Juan Feng. 2019. A causal inference method for reducing gender bias in word embedding relations.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2020. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology.

A Appendix



text string · lengths 0 149k	timestamp string · lengths 20 20	url string · lengths 14 811	clean bool 2 classes
This 2015 Volkswagen Golf R 2.0 Tsi Bluemotion 4motion – Not ed vehicle is located in Liverpool, L4 1RJ and the eBay seller is stesm_2776.	2019-04-21T18:15:56Z	https://www.accidentdamagedcars.org.uk/volkswagen/2015-volkswagen-golf-r-2-0-tsi-bluemotion-4motion-not-ed-5/	false
One of the prerequisites of staying in the Google Lunar XPRIZE (GLXP) is securing a contract with a launch service provider. India-based Team…	2019-04-24T12:51:34Z	https://www.spaceflightinsider.com/missions/commercial/team-indus-joins-google-lunar-x-prize-finalists-astrobotic-drops/	true
Julian is now on annual vacation until Tuesday 23 April. If you need to contact Julian during that period, please use email in the first…	2019-04-18T23:12:43Z	https://www.kw-law.co.uk/	true
Go to... Go to...	2019-04-20T11:14:48Z	http://barnhard.nl/2017/06/04/science_proves_that_homophobia_means_youre_probably_gay/	false
The Economist: "GREECE'S prime minister, George Papandreou, faced the television cameras on Friday 23rd April to announce that his government…	2019-04-20T06:14:09Z	http://radicalroyalist.blogspot.com/2010/04/hellenic-republic-is-losing-its.html	false
Please enable cookies. Why do I have to complete a CAPTCHA? Completing the CAPTCHA proves you are a human and gives you temporary access to th…	2019-04-19T22:24:53Z	http://duhocdaiviet.com/watch/skate-kitchen.html	true
Glen founded our firm in January 1978. Glen works in all areas of accounting and tax for businesses and individuals. His experience…	2019-04-22T20:54:40Z	https://milkustriercpa.com/our-team/glen-a-milkus-cpa/	true
Runway-inspired makeup to bust out for your next party. If you're anything like us, you've had your eye on the high-pigment makeup trends…	2019-04-24T18:22:46Z	https://blog.nastygal.com/beauty/2018/11/the-pigmented-eyeshadow-tutorial-you-need-now/	true

Figure 18: Common Crawl Dataset