Adaptive Reinforcement Learning for Mitigating Gender Bias in Natural Language Processing

Taahaa Mir
McGill University
taahaa.mir@mail.mcgill.ca

Abstract

In this paper, we introduce a novel reinforcement learning (RL) framework implemented in a custom Gym environment for adaptive gender debiasing of word embeddings. In our approach, **EmbeddingDebiasingEnv**, enables an RL agent to interact dynamically with embedding spaces, focusing on reducing gender bias while maintaining semantic integrity. The framework supports three actions: soft debiasing, counterfactual data augmentation (CDA), and a "Do nothing" option, which learns the utility of inaction in preserving optimal embeddings.

Our approach dynamically normalizes changes in bias and semantic similarity based on observed data, enabling the agent to adapt effectively across various embeddings. The reward function is designed to carefully balance bias reduction against semantic loss, with adjustments for each action's impact. We evaluate the agent's performance in multiple settings, demonstrating that it effectively reduces bias with minimal semantic degradation, with a goal to outperform traditional static methods. Preliminary results indicate our model's significant potential in reducing gender bias. This framework not only leverages the strengths of other debiasing techniques but also sets a foundation for addressing other biases in natural language processing.

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized numerous fields, including healthcare and finance, and have become entrenched in our daily lives. However, these systems often inherit human biases present in their training data[6], posing significant ethical challenges. These biases are particularly concerning as they can perpetuate harmful stereotypes and lead to unfair decision-making processes in critical areas such as employment, legal sentencing, and loan approvals.

Central to many NLP systems are word embeddings, which are mathematical representations capturing the meanings of words based on their contexts within training datasets. [13], [3], [11], [6]. These embeddings are used in a wide range of applications, from sentiment analysis[7] to machine translation[10]. However, if the training data contains gender, racial, or other biases, these biases can be reflected in the word embeddings, leading to biased outcomes in downstream applications. For example, in automated recruitment, AI-driven CV screening tools that rely on biased embeddings could disadvantage minority groups[1], underscoring the need for equitable AI practices.

Given the widespread use of word embeddings in NLP and the potential for bias, it is crucial to develop methods for debiasing these embeddings. By doing so, we can ensure that AI systems produce fair and unbiased outcomes, contributing to a more equitable society. This project introduces a novel reinforcement learning approach to debias word embeddings, aiming to mitigate gender bias effectively while preserving their functional utility across various NLP tasks.

The innovation in our approach lies in its dynamic adaptability and the reward function, which is sensitive to the subtle changes in word associations within different contexts. By integrating this method with the Q-Learning and Deep Q-Networks algorithms, we demonstrate through our experiments that it is possible to quantifiably reduce bias in word embeddings, paving the way for more equitable AI systems.

1 Background

Projecting onto the Gender Direction

It is crucial in our application to measure gender bias. One approach to analyze bias involves projecting word embeddings onto a “gender direction.” This direction is a vector in the embedding space that represents the concept of gender, typically

Submitted to McGill University in Partial Fulfillment of Requirements for Final Project for Graduate Course COMP579: Reinforcement Learning. Do not distribute.

constructed by calculating the difference between the vector representations of gender-specific words, such as "he" and "she."

The projection of a word vector onto the gender direction quantifies how much the word's representation is aligned with this conceptual gender axis. Mathematically, the projection of a vector $\vec{v}$ onto a direction $\vec{d}$ is given by:

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

where $\cdot$ denotes the dot product, and $\|\vec{d}\|$ is the norm of the gender direction vector. This formula helps quantify the extent to which individual word embeddings may inherently reflect gender biases.

Calculating Cosine Similarity

To assess the similarity between vectors, particularly to measure how changes in embeddings affect their original meanings, cosine similarity is used. Cosine similarity measures the cosine of the angle between two vectors:

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

In the debiasing context, maintaining high cosine similarity between the original and modified embeddings ensures that the debiasing process preserves the linguistic utility of the embeddings.

Counterfactual Data Augmentation (CDA)

Counterfactual Data Augmentation is an approach to mitigate bias by augmenting the training data with examples where sensitive attributes are altered but the context remains the same. For instance, in a gender debiasing scenario, sentences from the training corpus are duplicated with gender pronouns and gender-associated words swapped. This method aims to balance the representation in the training data, reducing the model's learned bias.

To implement CDA, our approach was:

1. Find a dataset with counterfactual examples [8] (Appendix: 7) where explicit gender indicators in texts are swapped.
2. Train a model, such as FastText [4], on this augmented dataset to obtain embeddings that potentially exhibit reduced bias.

Obtaining and Training on Common Crawl Data

To derive word embeddings that are both robust and contemporary, training on large, diverse datasets such as those provided by Common Crawl is advantageous. Common Crawl offers a broad snapshot of the internet, which includes texts from a multitude of domains and contexts. By training embedding models like FastText on data from Common Crawl, we can capture rich and varied semantic information.

The process involves:

1. Accessing Common Crawl data from Hugging Face (Appendix 6) [9].
2. Training the FastText model [4] on this data to produce word embeddings that are then used for analysis and debiasing tasks.

These preliminary concepts and methodologies form the foundation of our project, enabling a structured approach to understanding and mitigating gender bias in word embeddings.

2 Related Work

The field of mitigating bias in AI systems, particularly in word embeddings, has seen significant advancements in recent years. This section provides a brief overview of the literature in this field, focusing on the detection of bias in pre-trained embeddings and various mitigation techniques.

2.1 Bias in Pre-trained Embeddings

An empirical study by Sesari et al.[12] evaluated the bias of 15 publicly available, pre-trained word embeddings models based on three training algorithms (GloVe, word2vec, and fastText) with regard to four bias metrics (WEAT, SEMBIAS, DIRECT BIAS, and ECT). The study found that fastText was the least biased model in 8 out of 12 cases and that small vector lengths led to a higher bias[4].

2.2 Mitigation Techniques

Several mitigation techniques have been proposed to address the issue of bias in word embeddings. One such technique is soft debiasing, introduced by Bolukbasi et al. in their paper "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings"[5]. The authors proposed a method for modifying an embedding to remove gender stereotypes while maintaining desired associations.

Another technique is Counterfactual Data Augmentation (CDA), which was explored by Zmigrod et al. in their paper "Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology"[14]. The authors presented a novel approach for converting between masculine-inflected and feminine-inflected sentences in languages with rich morphology to reduce gender bias.

2.3 Reinforcement Learning for Bias Mitigation

Building upon these techniques, recent work has suggested the use of reinforcement learning algorithms for bias mitigation. A blog post by Victor Ashioya[2] proposed the use of Reinforcement Learning with AI Feedback (RLAIF) for mitigating gender bias in AI systems. The author suggested that RLAIF, along with other reinforcement learning algorithms such as RLHF (Reinforcement Learning with Human Feedback) and Factually Augmented RLHF, offer a data-driven, robust, and versatile approach to mitigating gender bias in AI systems.

Our work is based on these advancements and aims to further explore the use of reinforcement learning algorithms, specifically RLAIF, in debiasing word embeddings. By leveraging the strengths of these techniques, we aim to contribute to the ongoing efforts to create fair and equitable AI systems.

3 Methodology

3.1 Our Approach A.1:

In our study, we have developed a custom environment using OpenAI Gym. This environment simulates the scenario of embedding debiasing, allowing us to employ a Deep Q-Network (DQN) algorithm to learn effective strategies for reducing bias. The following sections detail the components of our reinforcement learning setup.

3.1.1 State Space

The state of our environment is defined by the following components:

- **Word Embedding:** The vector representation of the current word being processed.
- **Bias Vector:** A replicated vector where each element is the calculated bias of the current word embedding.
- **Semantic Similarity Vector:** A replicated vector representing the semantic similarity of the current word to its original embedding.
- **Action Vector:** A replicated vector of the last action taken, facilitating the tracking of previous actions' impacts.

3.1.2 Actions

The agent in our environment can take one of the following three actions, each intended to adjust the embedding towards reducing bias while attempting to preserve semantic integrity:

1. **Soft Debiasing:** Modifies the embedding to diminish gender bias. This method is effective for general debiasing but risks over-correcting gender-specific terms.
2. **Counterfactual Data Augmentation (CDA):** Adjusts embeddings by aligning them closer to a counterfactually augmented version, which helps in maintaining the meaning and nuances of the original text but might inherit biases from the training data.
3. **Do Nothing:** Maintains the current state of the embedding. This action is used to test the hypothesis that some embeddings are already optimally debiased and do not require further adjustment.

3.1.3 Reward Function

The reward function is formulated to quantitatively assess the effects of the agent's actions, defined in a piecewise manner:

$$\text{Reward} = \begin{cases} \gamma & \text{if action = `Do Nothing' and state is near-optimal} \\ -\gamma & \text{if action = `Do Nothing' and adjustment is needed} \\ \alpha \times (\text{Bias Reduction}) - \beta \times (\text{Semantic Change}) & \text{otherwise} \end{cases}$$

where:

- **Bias Reduction (Normalized)** is calculated as the difference in bias before and after the action, normalized over the observed range of bias changes.

- **Semantic Change (Normalized)** measures the change in semantic similarity, penalizing significant deviations from the original meaning.
- α **and** β are weights set empirically to balance bias reduction and semantic integrity.
- γ is the penalty or reward of taking no action depending on the state.

Measurement of Bias: Bias is quantified by projecting the word vector onto a predefined gender direction vector and calculating the magnitude of this projection.

Measurement of Semantic Similarity: Semantic similarity between two embeddings is measured using the cosine similarity, providing a scale from -1 (opposite meanings) to 1 (identical meanings).

Rationale for Actions: Including three distinct actions allows the model to utilize the strengths of both debiasing techniques while mitigating their weaknesses. Soft debiasing is adept at reducing overt biases but may inadvertently neutralize necessary gender-specific terms (See Figure 1). CDA preserves meaning better but risks incorporating biases present in the training data (See Figure 2). The 'Do Nothing' action assesses whether embeddings are already optimal, minimizing unnecessary modifications. This strategy ensures a nuanced approach to debiasing, dynamically adapted to each word's characteristics.

4 Experiments & Results

Hyperparameters:

Learning rate: 0.01, Exploration fraction: 0.1, Initial exploration probability: 1.0, Final exploration probability: 0.1, Total timesteps for training: 10,000, Number of episodes: 100, $\alpha = 0.6$, $\beta = 0.4$, $\gamma = 0.01$.

Benchmarks:

The benchmarks for the debiasing task are specifically derived from the performance metrics of the `EmbeddingDebiasingEnv` environment and the outcomes of the DQN agent's learning process:

- **Mann-Whitney U Test:** We employ the Mann-Whitney U test as a benchmark to statistically evaluate the effectiveness of the debiasing process. Specifically, we apply the test to compare the distributions of word projections along the gender direction before and after debiasing. The null hypothesis assumes that the distributions of both groups (pre- and post-debiasing) are equal. A statistically significant p-value (typically less than 0.05) indicates that the distributions differ significantly, signifying that the debiasing intervention had a measurable impact on reducing gender bias within the embeddings.
- **Graphical Visualizations for Bias Evaluation** We employ a series of graphical visualizations to evaluate the presence and extent of gender bias in word embeddings. The following visualizations are used:
  - **Cosine Similarity Plot:** We generate cosine similarity heatmaps to visually assess how closely words are associated with gendered terms. These plots provide a clear visual indication of the degree to which different words align with masculine or feminine vectors, highlighting potential areas of bias.
  - **Gender Projection Plot:** This plot illustrates the projection of professional titles and other words onto a predefined gender direction. By mapping these projections, we can visually identify which terms are more closely associated with male or female connotations within the embedding space, providing a direct visualization of gender bias.
  - **Boxplot of Projections:** We use boxplots to display the distribution of projections onto the gender direction for groups of words categorized by gender association. This type of plot is particularly useful for comparing the central tendencies and variabilities between male and female word groups, offering a straightforward comparative view of bias across different categories.

Results:

In our debiasing experiment, the cosine similarity heatmap (Appendix 5) revealed that the debiasing process was largely effective, with most words exhibiting projections close to zero, suggesting minimal gender bias. Words with inherent gender meanings, such as 'grandmother' and 'uncle', successfully retained their gender-specific associations, demonstrating the algorithm's capability to discern and preserve necessary gendered nuances. However, instances of over-debiasing were observed, as in the case of 'grandfather', which shifted towards a female association, indicating that while the algorithm generally reduces unwanted biases, it requires refinement. These results suggest that the current reward function needs to be optimized to maintain a balance between bias reduction and semantic integrity, underscoring the delicate nature of debiasing language models.

The boxplot visualization and the Mann-Whitney U test provide insightful evidence into the efficacy of our debiasing algorithm. Despite the statistical significance indicated by the Mann-Whitney U test p-value of 0.0079, the close proximity of the medians in the boxplot (Appendix 4) suggests that the central tendency of the gendered word embeddings' projections is indeed nearer to one another post-debiasing. This implies that while there is still a detectable difference between male and female word associations, the gap has been narrowed, indicating an improvement towards gender neutrality. The presence of outliers, as shown in the boxplot, highlights specific cases that require further attention, but the overall shift towards the center suggests our algorithm has made positive strides in reducing gender bias in the embeddings.

The observed reduction in bias is further supported by the direct comparison of word projections on the gender direction before and after the application of our debiasing algorithm (Appendix: 3). Notably, after debiasing, the projections are more closely aligned with the neutral midpoint, indicating a movement toward gender neutrality across the embedding space.

5 Conclusion and Future Work

The results of our debiasing experiment indicate a significant advancement towards gender-neutral word embeddings. As detailed in Appendix 5, the cosine similarity heatmap, and Appendix 3, our algorithm has minimized gender bias for a vast majority of words, aligning their projections closer to a neutral midpoint. Crucially, it has also maintained the semantic integrity of inherently gendered terms. Nevertheless, the emergence of over-debiasing in certain cases such as 'brother' suggests that further refinement of our algorithm is necessary. The statistically insignificant results from the Mann-Whitney U test highlight some issues which may demands further attention.

Looking ahead, future work will focus on fine-tuning the reward function to enhance the discrimination between necessary and arbitrary gender associations within embeddings. Moreover, it would be beneficial to expand this approach to other types of bias such as racial bias.

References

[1] Elham Albaroudi, Taha Mansouri, and Ali Alameer. "A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring". In: *AI* 5.1 (2024), pp. 383–404. ISSN: 2673-2688. DOI: 10.3390/ai5010019. URL: https://www.mdpi.com/2673-2688/5/1/19.

[2] Victor Ashioya. "Using reinforcement learning algorithms to mitigate gender bias". In: (2024).

[3] William Blacoe and Mirella Lapata. "A Comparison of Vector-based Representations for Semantic Composition". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Ed. by Jun'ichi Tsujii, James Henderson, and Marius Paşca. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 546–556. URL: https://aclanthology.org/D12-1050.

[4] Piotr Bojanowski et al. *Enriching Word Vectors with Subword Information*. 2017. arXiv: 1607.04606 [cs.CL].

[5] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings". In: *arXiv preprint arXiv:1607.06520* (2016).

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[7] Erion Çano and Maurizio Morisio. "Word Embeddings for Sentiment Analysis: A Comprehensive Empirical Survey". In: *CoRR* abs/1902.00753 (2019). arXiv: 1902.00753. URL: http://arxiv.org/abs/1902.00753.

[8] Anna Currey et al. *MT-GenEval: A Counterfactual and Contextual Dataset for Evaluating Gender Accuracy in Machine Translation*. 2022. arXiv: 2211.01355 [cs.CL].

[9] Keirp. *Common Crawl Sample Dataset*. Hugging Face Dataset Hub. 2021. URL: https://huggingface.co/datasets/keirp/common_crawl_sample.

[10] Basab Nath, Sunita Sarkar, and Narayan C. Debnath. "A Study of Word Embedding Models for Machine Translation of North Eastern Languages". In: *Computational Intelligence in Communications and Business Analytics*. Ed. by Kousik Dasgupta et al. Cham: Springer Nature Switzerland, 2024, pp. 343–359. ISBN: 978-3-031-48879-5.

[11] Tobias Schnabel et al. "Evaluation methods for unsupervised word embeddings". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 298–307. DOI: 10.18653/v1/D15-1036. URL: https://aclanthology.org/D15-1036.

[12] Emeralda Sesari, Max Hort, and Federica Sarro. "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)*. Ed. by Christian Hardmeier et al. Seattle, Washington: Association for Computational Linguistics, July 2022, pp. 129–144. DOI: 10.18653/v1/2022.gebnlp-1.15. URL: https://aclanthology.org/2022.gebnlp-1.15.

[13] Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. "Word Representations: A Simple and General Method for Semi-Supervised Learning". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Ed. by Jan Hajič et al. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 384–394. URL: https://aclanthology.org/P10-1040.

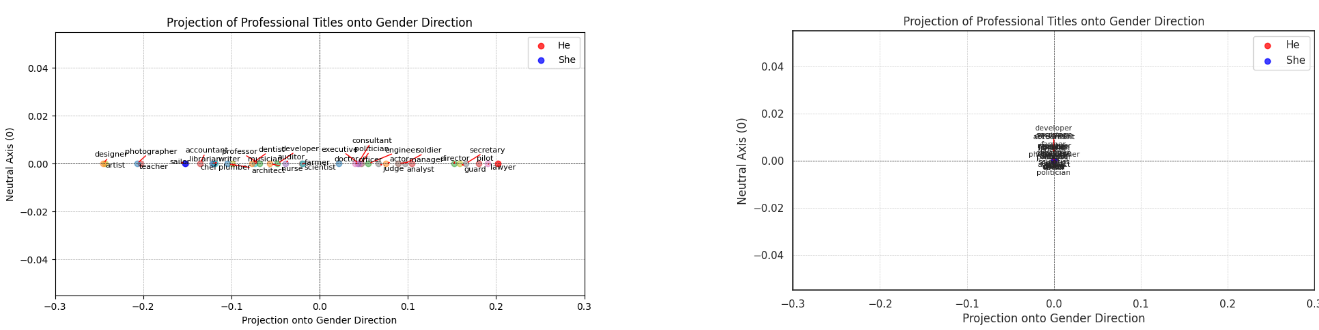
[14] Ran Zmigrod et al. *Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology*. 2020. arXiv: 1906.04571 [cs.CL].

A Appendix

A.1 Algorithm: Final Approach

```
Algorithm: Embedding Debiasing Environment
Input: embeddings, cda_embeddings, gender_direction, words_of_interest
Output: Debiasing of word embeddings through Reinforcement Learning

Initialize Environment:
    - Load word embeddings and counterfactual data augmentation (CDA) embeddings
    - Define gender direction and list of words of interest
    - Define action and observation spaces

Procedure Reset:
    - Randomly shuffle words of interest
    - Initialize current word and its embedding
    - Calculate initial bias and semantic similarity
    - Set initial state combining current embedding, bias, and semantic similarity

Procedure Check_Done_Condition:
    - Calculate cosine similarity for each word
    - If average similarity exceeds threshold or maximum iterations reached, return True

Procedure Step (action):
    - Apply action to current word embedding:
      - Soft Debiasing: Adjust embedding to reduce gender bias
      - CDA: Modify embedding towards CDA version
      - Do Nothing: Leave embedding unchanged
    - Update and normalize bias and semantic similarity changes
    - Calculate reward based on changes and predefined weights
    - Update environment state
    - Check if the environment meets done conditions

Main:
    - Initialize DQN agent with environment
    - Train DQN agent over specified number of episodes
    - Track and visualize episode rewards
```

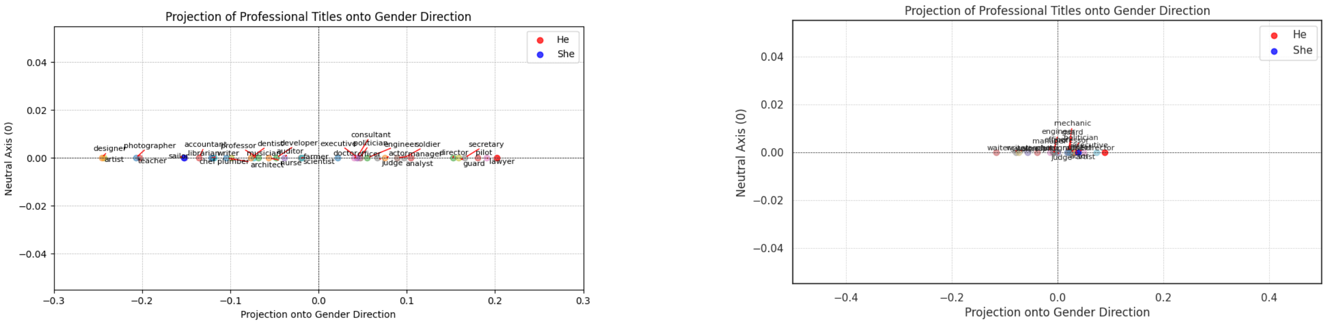
A.2 Figures



(a) Original Embedding

(b) Embedding after Soft Debiasing

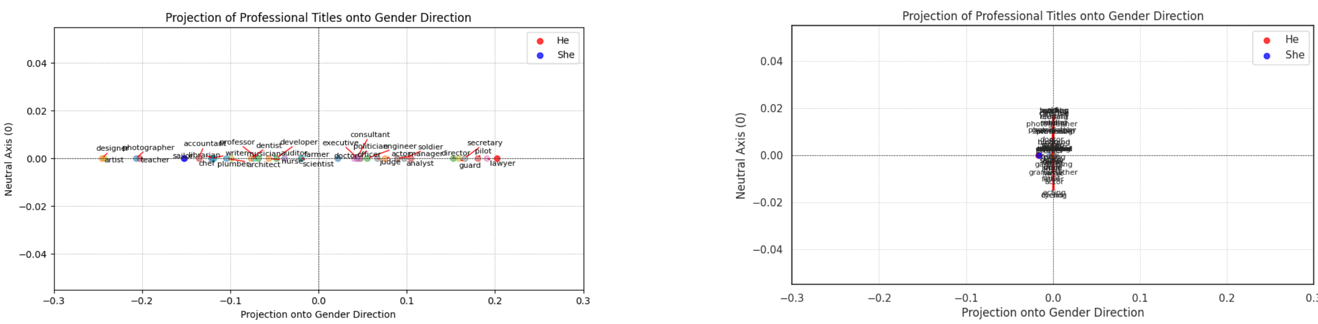
Figure 1: Original Embedding vs Embedding after Soft Debiasing



(a) Original Embedding

(b) Embedding after CDA

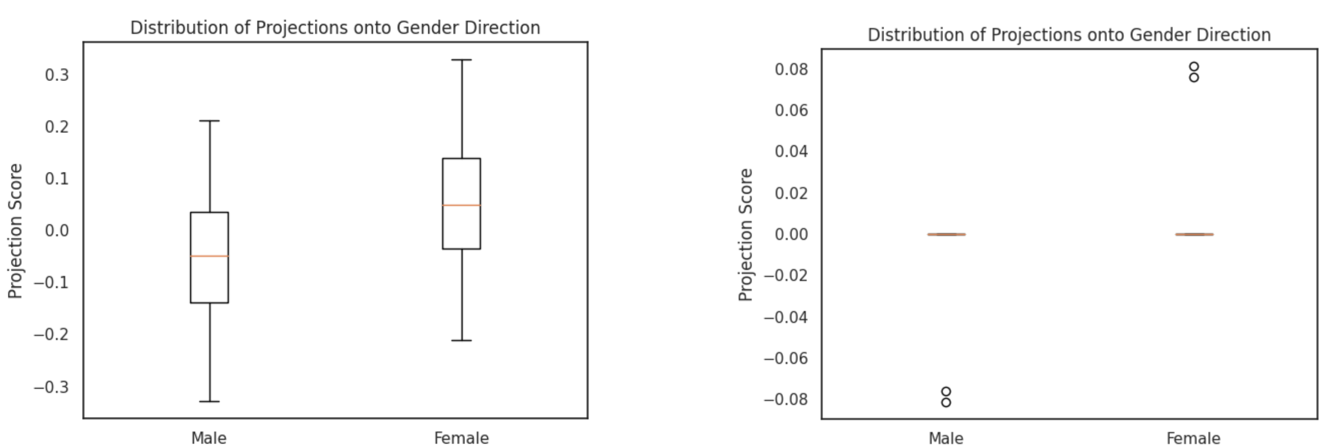
Figure 2: Original Embedding vs Embedding after CDA



(a) Embeddings before our Debiasing

(b) Embeddings after our algorithm

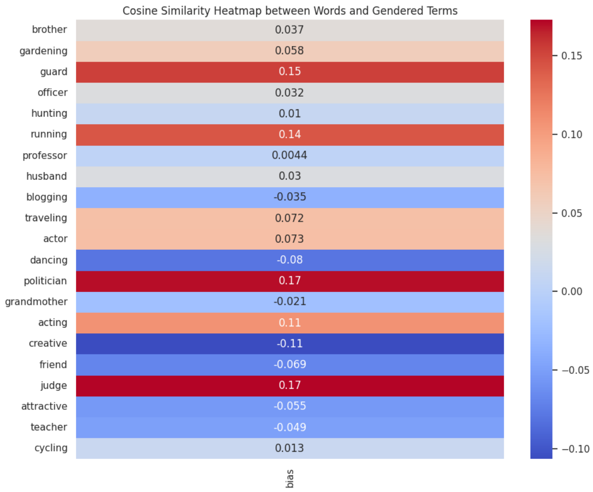
Figure 3: Original Embedding vs Embedding after our RL



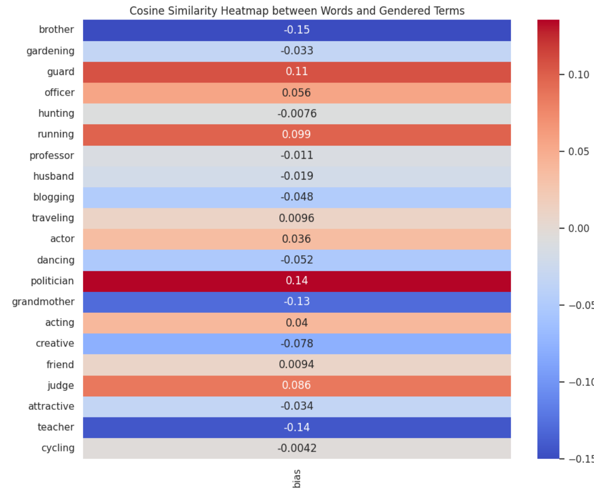
(a) Before our RL Debiasing

(b) After our RL Debiasing

Figure 4: Gender Projection Boxplot Before vs After our RL Algorithm



(a) Cosine Similarities with Gender Direction before RL



(b) Cosine Similarities with Gender Direction after RL

Figure 5: Cosine Similarities with Gender Direction before vs after RL

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 6: Common Crawl Dataset

```
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 7: Counterfactual Data Augmentation Dataset