Mitigating Bias in Word Embeddings: A Dual Approach Using Adversarial Debiasing and Hard-Debiasing

Ritik Rawat
Manan Davawala

Abstract

Word embeddings, such as those learned from models like GloVe and Word2Vec, have become a cornerstone in natural language processing (NLP), effectively representing words in a high-dimensional space where words with similar meanings are closer together. However, these embeddings often contain significant societal biases, including gender, racial, and occupational biases. This report presents a dual approach for mitigating gender bias in word embeddings: **Adversarial Debiasing** and **Hard-Debiasing**. Adversarial debiasing is a method where a predictor network is trained to learn embeddings while an adversary network simultaneously attempts to predict gender-related information. The goal is to neutralize these predictions, effectively “debiasing” the embeddings. Hard-debiasing, on the other hand, involves directly manipulating the word embeddings by projecting them onto a gender-neutral subspace. The results from these two approaches demonstrate significant reductions in gender bias in word embeddings, without sacrificing the semantic usefulness of the embeddings.

---

1. Introduction

1.1. Background

Word embeddings have transformed how machines understand and represent words in NLP. Algorithms like **GloVe** (Global Vectors for Word Representation) and **Word2Vec** have been trained on vast corpora of text to learn dense vector representations of words. These embeddings can capture a variety of syntactic and semantic relationships between words. For example, the vector difference between the words "king" and "queen" is very similar to the difference between "man" and "woman". However, recent studies have shown that these embeddings also inherit various societal biases embedded in the data used to train them.

One of the most concerning types of bias is **gender bias**. For instance, words like "nurse" and "teacher" are often more closely associated with female-related words, while "doctor" and "engineer" are associated with male-related words. These biases can perpetuate stereotypes and cause real-world harms, particularly when NLP models are used in decision-making processes such as hiring, healthcare, and criminal justice.

In this project, we focus on two primary methods for mitigating bias in word embeddings:

1. **Adversarial Debiasing**, where a model is trained to neutralize gendered information by minimizing the adversary's ability to predict gender-related attributes.
2. **Hard-Debiasing**, where gender-related components are directly removed from the embeddings using techniques like Principal Component Analysis (PCA) to adjust the vectors.

1.2. Problem Statement

The problem we aim to address is the **gender bias in word embeddings**. Given that these embeddings reflect harmful societal stereotypes, such as associating certain occupations with specific genders, it is crucial to develop methods to **de-bias** the embeddings while maintaining their ability to perform well in downstream NLP tasks. Failure to neutralize these biases can result in NLP models reinforcing these stereotypes in critical applications.

1.3. Objective

The objective of this report is to:

1. Investigate the effectiveness of **Adversarial Debiasing** and **Hard-Debiasing** as methods to mitigate gender bias in word embeddings.
2. Evaluate the quality of the resulting debiased embeddings using intrinsic and extrinsic evaluation metrics, ensuring that bias is reduced without sacrificing semantic quality or model performance on downstream tasks.

---

2. Related Work

The presence of bias in word embeddings has been widely discussed in the literature. One of the pioneering works in this domain was **Bolukbasi et al. (2016)**, who proposed the concept of **hard-debiasing**. Their method involved removing gendered dimensions from word embeddings by projecting the embeddings onto a subspace spanned by gendered terms, then neutralizing these components. This method was effective in reducing gender bias but came with the downside of losing some fine-grained semantic information.

Zhao et al. (2018) introduced **adversarial debiasing** as a technique to mitigate bias. The idea was to train a neural network with two components: a predictor that learns the word embeddings and an adversary that tries to predict gender. The key insight is that if the adversary cannot predict gender from the embeddings, the model must have successfully neutralized gender-specific information.

Other studies have explored **hybrid models** that combine these approaches, such as **Hard & Soft Debiasing** or adversarial debiasing followed by fine-tuning using supervised or

unsupervised learning techniques. These hybrid models are designed to remove bias while maintaining the utility of embeddings for various NLP tasks.

---

3. Methodology

3.1. Data and Word Embeddings

The core of our work is based on **GloVe embeddings**, which are pretrained word vectors obtained from a large corpus of text. These embeddings represent words in a 300-dimensional space. For the debiasing task, we use both gendered word pairs and occupation-related words. Gendered words, such as “he”, “she”, “man”, and “woman”, are used to compute the gender direction in the embedding space. Occupation words like “doctor”, “nurse”, “engineer”, and “teacher” are used to evaluate the effectiveness of debiasing.

3.2. Adversarial Debiasing

The adversarial debiasing method involves training a **generator model** and an **adversary model** in a manner that is similar to a **game-theoretic** setting, where the two models are adversaries in a **minimax** game. The key idea is that the predictor learns to generate embeddings, while the adversary attempts to predict gender from the embeddings.

1. **Generator Model**
   The predictor network takes the word embedding as input and generates a reconstructed version of the word embedding. The model architecture consists of 4 linear layers with ReLU activation along with a residual connection from the input to the output layer.

2. **Adversary Model**
   The adversary network attempts to predict the gender (male or female) of a word embedding. The goal of the adversary is to correctly classify whether the word is associated with a male or female gender. The model is a simple 3 layer binary classification model.

3. **Loss Functions**

   - **Generator Loss:** The predictor minimizes the reconstruction error (MSE), making it focus on learning word embeddings that preserve the meaning and syntactic properties of words.
   - **Adversary Loss:** The adversary learns to predict the gender of a word embedding using BCE loss.

4. **Optimization:**
   The training of this network is done in two phases. First phase involves training the predictor model independently to produce the original embeddings. During the second phase of the training, the two networks are trained alternately, with the adversary's weights updated to improve its ability to predict gender. However, the gradient reversal layer (GRL) is used to invert the gradients from the adversary's loss during backpropagation. This forces the predictor to **minimize the ability of the adversary to predict gender**, thus encouraging the embeddings to be neutral with respect to gender. This process ensures that the embeddings generated by the predictor contain as little gender information as possible. The label provided for the training are the direct dias scores. (Discussed Later in this report)

3.3. Hard-Debiasing

The hard-debiasing technique involves the following steps:

1. **Bias Direction Computation:**
   To identify the gender direction in the embedding space, we use **Principal Component Analysis (PCA)**. We collect a set of 10 paired gendered words (e.g., (‘he’, ‘she’),(‘man’, ‘woman’)) and compute their axis (man - woman). The principal component corresponding to the gender direction is computed, capturing the major axis of variation related to gender in the embedding space. This vector represents the "gender axis" — the direction in which gender-related differences between words exist in the word embedding space.

2. **Neutralizing Gendered Embeddings:**
   Once the gender axis is computed, we neutralize gendered embeddings by projecting the embeddings of gendered words (e.g., “he”, “she”, “doctor”, “nurse”) onto a subspace orthogonal to the gender axis. This is done by subtracting the projection of the word embeddings onto the gender axis. The projection is calculated using the dot product between the word embedding and the gender axis. Specifically, for each word embedding **v**:

$$v\_neutralized = \mathbf{v} - (\mathbf{v} \cdot \mathbf{g})\,\mathbf{g}$$

   Where:

   - **v** is the word embedding of the word (e.g., "doctor", "he").
   - **g** is the gender axis.
   - The term **(v · g)** gives the component of the word’s embedding along the gender direction. By subtracting this projection, we neutralize the word's embedding with respect to gender.

3. **Equalizing Gender-Neutral Embeddings:**

   After neutralizing the embeddings, we normalize them to ensure they retain their original scale and direction in other dimensions. This normalization helps maintain the original meaning and relationships of words while removing the gender-related component:

$$\mathbf{v\_neutralized} = \frac{\mathbf{v\_neutralized}}{\|\mathbf{v\_neutralized}\|}$$

   This ensures that the resulting embeddings are not biased along the gender axis but still retain their semantic properties.

---

4. Experiments and Results

4.1. Experimental Setup

To evaluate the effectiveness of our proposed debiasing methods, we performed a series of experiments using the following setup:

1. **Data:**

   We used pretrained **GloVe embeddings** (300-dimensional vectors), trained on a large corpus of text. We selected a set of gendered words (e.g., "he", "she", "doctor", "nurse") and occupation words (e.g., "engineer", "teacher", "doctor") as the main subjects of our experiments.

2. **Debiasing Methods:**

   We tested the following two debiasing methods:

   - **Adversarial Debiasing**: We implemented a neural network with a predictor and an adversary to remove gender bias from the embeddings by training them in an adversarial manner.

   - **Hard-Debiasing**: We applied PCA to compute the gender axis, followed by subtracting the projection of word embeddings onto this axis to neutralize gender bias.

3. **Evaluation Metrics:**

   - **Direct Bias**: We used the mean of **cosine similarity** between gender-neutral words and the gender axis as a measure of bias.

- **Semantic Quality**: We measured the quality of the embeddings by evaluating the **word similarity** and **analogy** tasks using the **WordSim-353** and **RG-65** datasets. The effectiveness of debiasing was considered successful if there was a significant reduction in bias without significant loss in semantic relationships.

4.2. Results

The results of the experiments showed the following:

	Direct Bias	**RG-65**	**WordSim-353**
Original	0.0919	0.7317	0.6666
Hard Debiasing	0.0052	0.7167	0.6521
Adversarial Debiasing	1.25×10^{-8}	0.7242	0.6641

1. **Adversarial Debiasing -**
    - The adversarial debiasing method significantly reduced gender bias in the word embeddings as measured by direct bias.
    - The predictor's performance remained strong, as there was minimal loss in the ability of the embeddings to preserve semantic information.

2. **Hard-Debiasing**:
    - The hard-debiasing approach successfully neutralized nearly all of the gender bias associated with the embeddings.
    - While the embeddings were debiased, a slight loss in certain semantic relationships was observed.

---

5. Our Contribution

Our contribution focuses on advancing the adversarial debiasing approach through an innovative model architecture and optimization technique. We designed a generator with residual connections, enabling it to produce embeddings that remain as close as possible to the original embeddings while addressing bias. This architecture ensures that the transformed embeddings retain their essential characteristics. Additionally, we employed a Gradient Reversal Layer (GRL) as part of the optimization process, allowing gradients to flow back to the generator effectively. This setup ensures that the generator is trained in a way that minimizes bias while preserving the integrity of the original embeddings.

5. Discussion

While both debiasing methods show promising results, there are some challenges and considerations:

1. **Loss of Subtle Semantic Information**:
   One of the trade-offs of debiasing techniques, particularly **hard-debiasing**, is that while gender-related biases are mitigated, subtle nuances in word meanings can be lost. For example, removing gender associations from words like "nurse" or "teacher" might slightly alter the semantic relationship these words have with each other, as they are historically tied to certain genders. Although the embeddings retain overall utility in tasks like word similarity, it's important to carefully consider how debiasing may affect specific downstream applications, particularly those that rely on nuanced semantic relationships.

2. **Scalability to Other Types of Bias**:
   The methods used in this project are tailored towards **gender bias** in word embeddings. However, societal biases extend beyond gender, encompassing racial, ethnic, and socio-economic biases, among others. While the adversarial debiasing method is flexible and can, in theory, be extended to mitigate other biases, **hard-debiasing** requires new computations for each type of bias. Extending both approaches to handle other biases would require a broader exploration of the embedding space and how other societal factors manifest in word relationships.

3. **Potential Ethical Concerns**:
   While mitigating bias is an ethical necessity in NLP, it also presents challenges. For example, the hard-debiasing approach involves projecting embeddings onto a neutral subspace, if doing the same for racial biases, it might be seen as "erasing" certain cultural or historical aspects of language . This raises the question: **to what extent should we alter word representations?** The implications of making certain words neutral can be ethically contentious, especially when considering the cultural or social weight certain words carry.

4. **Adversarial Debiasing as a Long-Term Solution**:
   Adversarial debiasing demonstrates that we can achieve better results in terms of reducing bias without substantial loss of performance. However, it is important to note that adversarial training can be computationally expensive and might not be suitable for all applications, especially when deploying at scale. Moreover, adversarial training depends on the presence of a strong adversary and may require fine-tuning over many epochs to stabilize.

---

6. Conclusion

This project demonstrates the effectiveness of **adversarial debiasing** and **hard-debiasing** techniques in mitigating gender bias within word embeddings, such as those generated by GloVe. Both methods, when applied individually, significantly reduce gendered associations in word representations without compromising the ability of the embeddings to preserve semantic relationships. However, as we observed, both methods have their limitations, including potential loss of subtle semantic relationships and challenges in extending the debiasing process to other types of societal biases.

Adversarial debiasing is effective in preventing the model from learning gendered representations by training the predictor and adversary networks in a minimax framework. This technique successfully reduces bias while retaining semantic usefulness. **Hard-debiasing**, on the other hand, neutralizes gendered directions in the embedding space, providing a geometric method for bias reduction. While effective, it may result in the loss of some contextual information and may be more difficult to apply to biases beyond gender.

In future work, it would be valuable to explore more sophisticated hybrid models that combine adversarial debiasing with other techniques, such as **counterfactual data augmentation** or **bias-sensitive fine-tuning**. Additionally, extending the debiasing methods to handle multiple biases simultaneously (such as racial or occupational bias) would be an important step toward achieving fairer NLP models across all domains.

---

7. Future Work

7.1. Multidimensional Bias Mitigation

While this project focused on gender bias, future work should consider the mitigation of multiple types of bias in word embeddings. **Biases related to race, ethnicity, age, and socio-economic status** are prevalent in NLP tasks, and a more generalized debiasing approach is needed. One possibility is to train adversarial networks that target different types of biases simultaneously, or to perform **multi-dimensional hard-debiasing** to neutralize several biases at once.

7.2. Task-Specific Debiasing

Another potential avenue is to tailor the debiasing process based on specific tasks. While debiasing is generally beneficial, some tasks (e.g., sentiment analysis or text classification) might be more sensitive to biases and require a higher degree of debiasing. Future work could involve **task-specific debiasing frameworks** that adjust

the level of debiasing applied to embeddings based on the nature of the downstream task.

7.3. Fine-Grained Evaluation

Our evaluation metrics focused on measuring the **reduction of bias** and the **preservation of semantic information**, but future work could expand on these to include more fine-grained analyses. This might include:

- Measuring bias across different linguistic features (e.g., pronouns, adjectives, occupations).
- Evaluating the impact of debiasing on different types of word relationships (e.g., synonyms, antonyms, analogies).
- Incorporating **human evaluation** to assess how well debiased embeddings align with real-world societal norms and values.

---

8. References

1. **Bolukbasi, T., Chang, W. W., Zou, J. Y., Saligrama, V., & Kalai, T. T. (2016)**. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv preprint arXiv:1607.06520*.
2. **Zhao, J., Chang, K. W., Yatskar, M., Ordonez, V., & Chang, W. (2018)**. Gender Bias in Contextualized Word Embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1-6.
3. **Peters, M. E., et al. (2018)**. Deep Contextualized Word Representations. *Proceedings of NAACL-HLT 2018*.
4. **GloVe: Global Vectors for Word Representation** (2014). *Stanford NLP Group*.
