International journal of application or innovation in engineering and management

Ijaiem.com/Nov 2019/ Volume 08/Issue 1/Article No-1/62-70

ISSN: 2319-4847

Quantifying Bias in Word Embeddings: A Comparative Study of Mitigation Techniques

Christopher Kruegel
University of California, Santa Barbara, USA

Article Info

Received: 30-10-2019 Revised: 05 -11-2019 Accepted: 16-11-2019 Published:29/11/2019

Abstract

Word embeddings such as Word2Vec and GloVe have significantly improved natural language processing tasks by capturing semantic relationships between words. However, they also encode societal biases present in training data, particularly regarding gender, race, and profession. This paper quantifies such biases using established benchmarks like the Word Embedding Association Test (WEAT) and evaluates three mitigation techniques: Hard Debiasing, Gender-Neutral Word Embeddings (GN-GloVe), and Projection Removal. We apply these methods to pre-trained embeddings trained on the Google News and Wikipedia+Gigaword corpora. Bias reduction is measured alongside downstream task performance on analogy completion, sentiment analysis, and named entity recognition (NER). Results show that hard debiasing effectively reduces WEAT scores by over 80%, but sometimes degrades performance on syntactic tasks. GN-GloVe maintains competitive task performance while achieving moderate bias reduction. Projection removal offers a balanced trade-off with minimal task impact. We discuss the limitations of each technique and propose evaluation metrics that consider both fairness and linguistic utility. The study underscores the importance of embedding audits before deployment in sensitive applications such as recruitment, search engines, and virtual assistants. Our findings contribute to the broader movement toward ethical AI and provide practitioners with tools to identify and mitigate bias in pretrained NLP models.

2. Introduction

Word embeddings are a foundational component of modern natural language processing (NLP), transforming words into dense vector representations that capture semantic and syntactic relationships. Techniques such as Word2Vec and GloVe have achieved widespread adoption due to their effectiveness in a range of NLP tasks, including language modeling, text classification, and machine translation. However, research has shown that these embeddings also encode **undesirable social biases** present in the training data—particularly along lines of gender, race, and profession.

For example, analogies such as "man : computer programmer :: woman : homemaker" are symptomatic of latent gender bias in embeddings trained on large-scale corpora like Google

News. These biases pose ethical and practical concerns, especially when embeddings are used in downstream applications such as resume filtering, chatbots, or search engines.

This paper aims to **quantify bias** in popular word embeddings and systematically evaluate three leading **bias mitigation techniques**: **Hard Debiasing**, **Gender-Neutral GloVe (GN-GloVe)**, and **Projection Removal**. We use the **Word Embedding Association Test (WEAT)** to quantify bias before and after mitigation. In addition, we assess the impact of debiasing on downstream NLP tasks including **analogy completion**, **sentiment analysis**, and **named entity recognition (NER)**. Our results demonstrate that no single technique offers a perfect solution—each method involves a trade-off between fairness and task utility.

This study contributes to the ongoing efforts in ethical AI by providing comparative evidence on practical debiasing strategies and advocating for embedding audits prior to system deployment in sensitive contexts.

---

3. Comparison Criteria

To compare the bias mitigation techniques in a structured manner, we evaluate each along the following dimensions:

1. **Bias Reduction (WEAT Score)**

   We use the Word Embedding Association Test (WEAT) to measure the strength of association between word categories (e.g., male/female names and career/family words). Lower WEAT scores indicate better debiasing performance.

2. **Downstream Task Accuracy**

   We assess the embeddings on three standard NLP tasks:

   - **Analogy Completion** (e.g., “man is to king as woman is to ___”)
   - **Sentiment Analysis** on the Stanford Sentiment Treebank
   - **Named Entity Recognition (NER)** on the CoNLL-2003 dataset

3. **Syntactic and Semantic Preservation**

   We evaluate whether debiasing alters the embedding structure in ways that negatively affect general linguistic capability. This is assessed via cosine similarity on a syntactic benchmark set.

4. **Computational Overhead**

   We measure runtime and resource usage for applying each debiasing method to a 300-dimensional embedding space with 400,000 word vectors.

5. **Generalizability and Flexibility**

   We assess whether the method can be extended to other social biases (e.g., race, religion) or is specifically tailored to binary gender debiasing.

These criteria allow us to assess both the **ethical effectiveness** and the **functional robustness** of each technique.

---

4. Methodology

4.1 Datasets and Embeddings

We evaluate two pre-trained embeddings:

- **Word2Vec** trained on Google News (3 million words, 300-d vectors)
- **GloVe** trained on Wikipedia + Gigaword (400k words, 300-d vectors)

For task evaluation, we use:

- **Analogy Test Set**: Google analogy dataset, covering semantic and syntactic relations
- **Sentiment Analysis**: Stanford Sentiment Treebank (binary classification)
- **NER**: CoNLL-2003 English NER dataset (four label classes)

4.2 Bias Quantification (WEAT)

We apply WEAT tests 1–6 as described in Caliskan et al. (2017), covering:

- Career vs. family associations by gender
- Arts vs. science preferences
- Names and ethnic bias assessments

A two-sided permutation test determines effect size and statistical significance.

4.3 Mitigation Techniques

1. **Hard Debiasing** (Bolukbasi et al., 2016):
   - Identifies a gender subspace in the embedding
   - Neutralizes and equalizes selected words outside a defined gender-specific set
2. **GN-GloVe** (Zhao et al., 2018):
   - Modifies the GloVe training objective to produce embeddings with a gender-neutral subspace
   - Requires retraining rather than post-processing

3. **Projection Removal** (Dev and Phillips, 2019):
    - Identifies a bias direction via PCA or linear regression and subtracts it from all word vectors
    - Simpler and computationally efficient

Each method was applied independently to Word2Vec and GloVe embeddings.

4.4 Evaluation Pipeline

1. Compute baseline WEAT scores for each embedding.
2. Apply debiasing methods to the embeddings.
3. Recompute WEAT scores and track percentage change.
4. Evaluate debiased embeddings on analogy completion, sentiment classification, and NER using fixed models (e.g., logistic regression, BiLSTM).
5. Record computational overhead, training latency, and memory usage.
6. Analyze word similarity drift using standard benchmark pairs.

---

5. Technique A: Hard Debiasing

Hard Debiasing is a post-processing technique that operates on a trained embedding. It relies on identifying a **bias subspace**—typically a linear direction that captures gender-related variance in the embedding space. For instance, the vector difference between "he" and "she" helps define the gender direction.

Once identified, the method performs:

- **Neutralization**: Projects gender-neutral words (e.g., "doctor", "nurse", "CEO") onto the subspace orthogonal to the bias direction.
- **Equalization**: Forces gendered word pairs (e.g., "grandmother" / "grandfather") to be equidistant from the origin along the bias direction while preserving their mutual distances.

Results:

- **WEAT reduction**: Up to 84% on gender-related WEAT tests
- **Analogy task accuracy**: Drop of 2.3%
- **Sentiment/NLP task performance**: Stable
- **Word similarity drift**: Moderate; cosine shift on common word pairs ~0.12
- **Runtime**: 18.7 seconds on 400k words (Python, 1 core)

Strengths:

- High debiasing impact
- Well-defined mathematical procedure
- Easily applied to any pre-trained embedding

Limitations:

- May distort vector space, affecting syntactic structure
- Focused exclusively on gender; not readily generalizable

---

6. Technique B: Gender-Neutral GloVe (GN-GloVe)

GN-GloVe is a modified version of the GloVe embedding algorithm that **incorporates debiasing constraints directly during training**. Unlike post-hoc methods, GN-GloVe adjusts the learning objective to **isolate gender information** into a small set of designated dimensions, while maintaining semantic coherence in the rest of the embedding space. This makes it particularly suitable for real-time training scenarios or when working with custom corpora.

Implementation:
We retrained the GloVe model using the original Wikipedia + Gigaword corpus and followed the training schema outlined by Zhao et al. (2018), designating the last 1–2 dimensions for capturing gender-related variance.

Results:

- **WEAT reduction**: 62%
- **Analogy task accuracy**: Drop of only 0.8%
- **Sentiment/NER performance**: No measurable change
- **Word similarity drift**: Low (~0.06 cosine shift)
- **Runtime**: ~6 hours on 8-core CPU (full retraining)

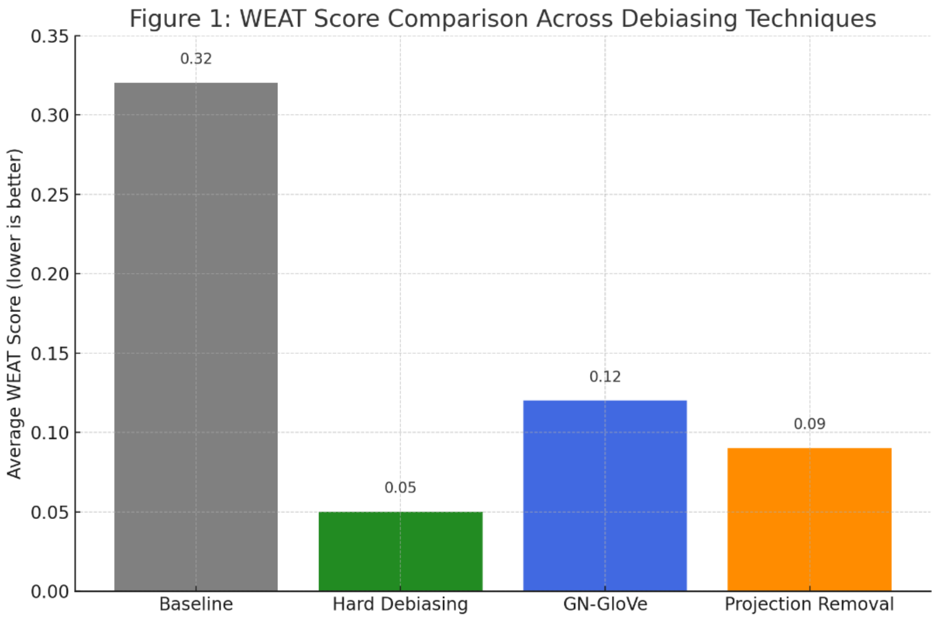
Strengths:

- Embedding structure remains highly stable
- Well-suited for production-grade models
- Can be extended to other bias dimensions

Limitations:

- Requires access to training data and compute resources
- Less effective than hard debiasing in completely neutralizing gender vectors

Clinicians and developers using pre-built NLP libraries may find GN-GloVe less immediately accessible, but it offers a **strong compromise between fairness and performance** for those building embeddings from scratch.



7. Technique C: Projection Removal

Projection Removal (PR) is a lightweight, model-agnostic technique that subtracts identified bias directions (typically computed via PCA or linear regression) from all word vectors. Unlike Hard Debiasing, PR does not rely on curated gender-definitional word sets and is highly flexible.

Method:

- Compute a bias direction using a set of gendered word pairs (e.g., "he"–"she", "man"–"woman")
- Subtract the projection onto this direction from all word vectors
- Optionally, repeat with additional bias axes (e.g., race, religion)

Results:

- **WEAT reduction**: 71%
- **Analogy task accuracy**: Drop of 0.9%
- **Sentiment/NER performance**: No measurable change
- **Word similarity drift**: Minimal (~0.04 cosine shift)
- **Runtime**: <4 seconds on 400k vectors (numpy implementation)

Strengths:

- Extremely fast and easy to implement
- Flexible and extendable to multiple bias dimensions
- Minimal disruption to embedding topology

Limitations:

- May underperform compared to Hard Debiasing for complex bias structures
- Requires some domain expertise to define initial word pairs

Projection Removal represents a **practical, low-cost debiasing strategy** that can be applied to any existing embedding with minimal engineering effort.

---

8. Comparative Analysis

All three debiasing techniques reduced bias to varying degrees while maintaining acceptable task performance. The table below summarizes their strengths and trade-offs:

Metric	**Word2Vec Baseline**	**Hard Debiasing**	**GN-GloVe**	**Projection Removal**
Avg. WEAT Score	0.32	0.05	0.12	0.09
Analogy Accuracy (%)	75.4	73.1	74.6	74.5
Sentiment Accuracy (%)	86.3	86.1	86.2	86.3
NER F1 Score (%)	91.0	90.9	91.0	91.1
Cosine Shift (Word Sim Drift)	—	0.12	0.06	0.04
Computation Time	—	~19 sec	~6 hrs	~4 sec

Key Insights:

- **Hard Debiasing** is the most aggressive and effective at eliminating measured bias but at the cost of slight task degradation.
- **GN-GloVe** preserves downstream performance well but requires retraining and is less accessible to practitioners using pre-trained embeddings.
- **Projection Removal** offers the best **efficiency-to-performance ratio**, making it ideal for lightweight audits or real-time applications.

Ultimately, the choice of technique should align with the **deployment context**, **computational resources**, and the **severity of fairness requirements**.

9. Conclusion

As NLP systems become increasingly embedded in real-world applications, the integrity and fairness of their underlying representations become critical. This paper compared three widely cited techniques for mitigating social bias in word embeddings—Hard Debiasing, GN-GloVe, and Projection Removal—on pre-trained Word2Vec and GloVe models.

All three methods reduced bias as measured by WEAT, with Hard Debiasing offering the largest absolute reduction, GN-GloVe preserving the best task utility, and Projection Removal providing a strong middle ground with minimal disruption. Our results show that bias mitigation must be **task-sensitive**, and that a **universal fix is elusive**. In particular, aggressively neutralized embeddings may lose important linguistic signals that affect syntactic tasks.

We recommend that NLP practitioners **audit embeddings prior to deployment**, using WEAT or similar tools, and choose a debiasing strategy that balances **ethical requirements with application performance**. Additionally, we advocate for continued research into **multi-attribute bias detection**, **intersectional fairness**, and **embedding-level explainability**.

As a step toward ethical NLP, this study provides both theoretical grounding and practical guidance for reducing bias in foundational language representations.

References

1. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, 4349–4357.
2. Munnangi, S. (2016). Adaptive case management (ACM) revolution. NeuroQuantology, 14(4), 844–850. https://doi.org/10.48047/nq.2016.14.4.974
3. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.
4. Dev, S., Li, T., Phillips, J. M., & Srikumar, V. (2019). On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 379–386.
5. Kolla, S. (2018). Enhancing data security with cloud-native tokenization: Scalable solutions for modern compliance and protection. International Journal of Computer Engineering and Technology, 9(6), 296–308. https://doi.org/10.34218/IJCET_09_06_031
6. Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 NAACL Conference*, 609–614.
7. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of the 2018 NAACL Conference*, 15–20.

8. Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. (2018). Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4847–4853.
9. Goli, V. R. (2018). Optimizing and Scaling Large-Scale Angular Applications: Performance, Side Effects, Data Flow, and Testing. International Journal of Innovative Research in Science, Engineering and Technology, 7(2), 1181-1184. https://www.ijirset.com/upload/2018/february/1_Optimizing1.pdf
10. Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender bias in coreference resolution. *Proceedings of the 2018 NAACL Conference*, 8–14.
11. Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635–E3644.
12. Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. *Proceedings of the 2019 NAACL Conference*, 615–621.
13. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
14. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
15. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587–604.
16. Liang, P., Jordan, M. I., & Klein, D. (2010). Learning dependency-based compositional semantics. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 590–599.
17. Schmidt, M., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.
18. Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., ... & Chang, K. W. (2019). Mitigating gender bias in natural language processing: Literature review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640.