How to quantify bias in word embedding : WEAT

Ga-Hyun Oh
lucy071101@gmail.com

June 4, 2025

Abstract

Bias is one of the most prominent issues in AI ethics. The first step toward addressing this issue is to quantify it. To this end, we use the Word Embedding Association Test (WEAT), which is based on cosine similarity. WEAT allows us to measure bias in embeddings and visualize it using heatmaps and principal component analysis (PCA). Furthermore, by identifying and quantifying biased associations in word embeddings, WEAT can play a key role in enhancing embedding fairness and fostering socially beneficial outcomes.

1 Introduction

One of the earliest and most symbolic incidents that brought widespread public attention to the risks of AI bias was Microsoft's chatbot Tay, released in 2016. This case demonstrated how biases embedded in algorithms and AI systems can lead to serious ethical issues. As highlighted in a journal article by Garcia [1], such risks underscore the need for multi-layered efforts to address algorithmic bias. In response to growing concerns, subsequent research has examined gender bias in word embeddings, particularly in GloVe models trained on Google News data [2]. Building on this line of work, we define WEAT to quantitatively assess and generalize such biases across various datasets.

2 Method

2.1 Dataset and Embedding Methodology

We used a Korean-language movie synopsis dataset, which was separated into two categories: art films and commercial (mainstream) films. Based on this separation, we constructed distinct word sets for each category. Additionally, we created word sets for each genre within the dataset. For morphological analysis, we employed the Okt (Open Korean Text) tokenizer. To train word embeddings, we used the Word2Vec model implemented in Gensim 4.0.

2.2 Preprocessing for Fair WEAT Scoring

To refine the input vocabulary for bias evaluation, we first represented the corpus as a sparse matrix using Term Frequency-Inverse Document Frequency (TF-IDF). In order to eliminate words that are commonly found in both art and commercial film synopses—regardless of category—we filtered out terms that appeared across both domains. This step was intended to remove genre-independent words that do not contribute meaningfully to the bias measurement. The resulting word sets were then used to calculate the WEAT scores.

2.3 WEAT score

a. Cosine Similarity

$$\cos\_\text{sim}(\mathbf{i}, \mathbf{j}) = \frac{\mathbf{i} \cdot \mathbf{j}}{\|\mathbf{i}\| \cdot \|\mathbf{j}\|} \tag{1}$$

Cosine similarity measures the cosine of the angle between two vectors $\mathbf{i}$ and $\mathbf{j}$. It is commonly used to evaluate semantic similarity between word embeddings, ranging from -1 (completely opposite) to 1 (identical direction).

b. Association Score for a Word

$$s(\mathbf{w}, A, B) = \frac{1}{|A|} \sum_{\mathbf{a} \in A} \cos(\mathbf{w}, \mathbf{a}) - \frac{1}{|B|} \sum_{\mathbf{b} \in B} \cos(\mathbf{w}, \mathbf{b}) \tag{2}$$

The association score $s(\mathbf{w}, A, B)$ quantifies how strongly a target word $\mathbf{w}$ is associated with two sets of attribute words, A and B. A higher score indicates a stronger association of $\mathbf{w}$ with A compared to B.

c. WEAT Score

$$\text{WEAT}(X, Y, A, B) = \frac{\mu_{x \in X} \left[s(x, A, B) \right] - \mu_{y \in Y} \left[s(y, A, B) \right]}{\sigma_{w \in X \cup Y} \left[s(w, A, B) \right]} \tag{3}$$

The Word Embedding Association Test (WEAT) score compares the relative association of two target word sets, X and Y, with two attribute sets, A and B. It is calculated by taking the difference in mean association scores and normalizing it by the pooled standard deviation. A higher absolute WEAT score indicates a stronger bias.

3 Result

Following the methodology described in the *Method* section, we calculated WEAT scores for pairs of genres to assess their relative closeness to mainstream or art films. To focus on meaningful distinctions, we selected genre pairs with WEAT scores less than or equal to –0.8 or greater than or equal to +0.8.

Table 1 lists genre pairs with WEAT scores ≤ -0.8, indicating that Genre A is relatively closer to mainstream films, while Genre B is relatively closer to art films.

Conversely, Table 2 shows genre pairs with WEAT scores $\geq +0.8$, where Genre A is relatively closer to art films, and Genre B is closer to mainstream films.

These results provide a clear separation between genres associated with mainstream and art cinema, as captured by the WEAT metric.

Genre A	**Genre B**	**WEAT Score**
Others	Drama	**–0.9821**
Documentary	Drama	**–0.9197**
Documentary	Adventure	**–0.9130**
Documentary	Comedy	**–0.8693**
Documentary	Romance	**–0.8603**
Others	Romance	**–0.8571**
Others	Comedy	**–0.8503**
Others	Adventure	**–0.8428**
Documentary	Historical	**–0.8258**

Table 1: Genre pairs with the lowest WEAT scores

Genre A	**Genre B**	**WEAT Score**
Performance	Musical	**+1.0718**
Adventure	Fantasy	**+1.0419**
Performance	Animation	**+0.9607**
Drama	Animation	**+0.8677**
Family	Animation	**+0.8557**
Family	Others	**+0.8641**
Family	Documentary	**+0.8524**
Comedy	Fantasy	**+0.8428**
Melodrama	Fantasy	**+0.8328**
Drama	Musical	**+0.8293**
Melodrama	Animation	**+0.8149**
Melodrama	Musical	**+0.8132**

Table 2: Genre pairs with the highest WEAT scores

To complement the quantitative analysis provided by the WEAT scores, we employed a heatmap (Figure 1) and Principal Component Analysis (PCA) visualization (Figure 2) for further exploration of genre relationships in the embedding space.

The heatmap (Figure 1) provides an intuitive overview of WEAT scores across multiple genre pairs, highlighting clusters or patterns of bias that might not be immediately apparent from individual scores alone.

PCA (Figure 2) reduces the high-dimensional embedding vectors into two principal components, allowing visualization of the relative positions of genres. This dimensionality reduction reveals the underlying structure and grouping of genres, making it easier to interpret how certain genres relate to mainstream or art films in the embedding space.

Together, these visualizations complement the numerical WEAT analysis by offering deeper insight into genre biases and relationships.

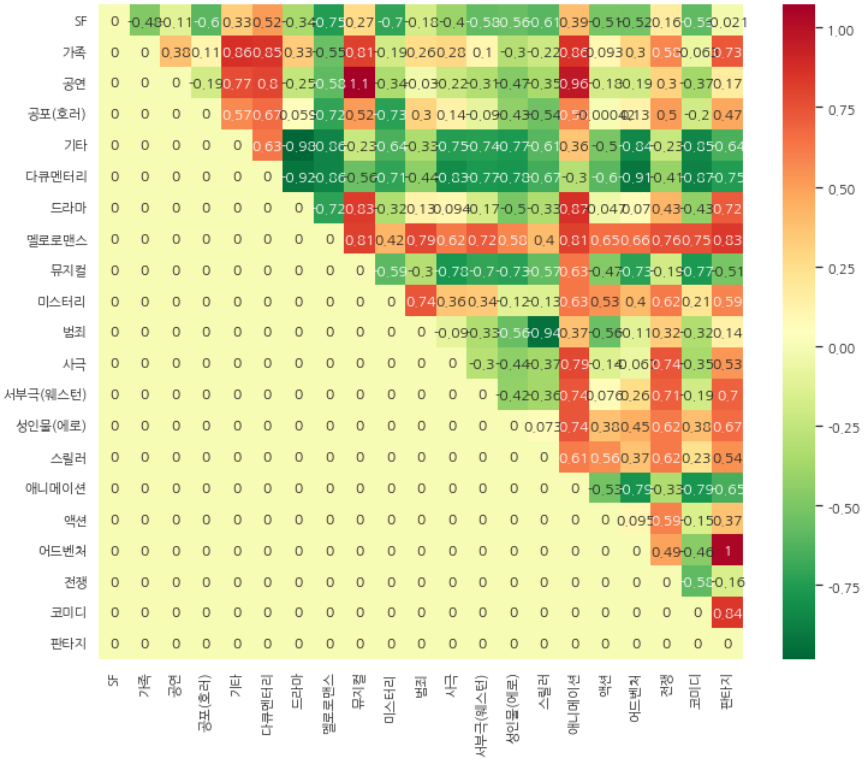

Figure 1: Heatmap showing WEAT scores across genre pairs

Figure 2: PCA visualization of genre embeddings in two dimensions

4 Conclusion

Beyond simply improving embedding fairness, this work holds significance as a potential exemplary approach for addressing broader ethical challenges in artificial intelligence. A key limitation of our study is that, while we quantified bias through the WEAT score, we did not implement techniques to mitigate such biases within models. Therefore, we anticipate that future research will build upon the WEAT metric to develop effective debiasing methods, which not only resolve embedding bias. Ultimately, we hope that the experience gained from addressing bias through these techniques will inform and inspire solutions to a wider range of ethical concerns in AI.

References

[1] M. GARCIA. "RACIST IN THE MACHINE: THE DISTURBING IMPLICATIONS OF ALGORITHMIC BIAS". In: *World Policy Journal* 33.4 (2016), pp. 111–117. ISSN: 07402775, 19360924.

[2] T. Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 29. 2016, pp. 4349–4357.