WEAT-Based Quantitative Analysis of Bias in Movie Corpus Word Embeddings: Focus on Art vs. General Films Across Genres

Jiyeon Seo
NLP Track, AIFFEL
Modulabs, South Korea
yjneon339@gmail.com

Abstract

This study investigates semantic bias in word embeddings related to movie classification (art vs. commercial films) using a movie synopsis corpus. Noun morphemes were extracted with MeCab, and CBow and Skip-gram models were trained. Target and attribute word sets were constructed using TF-IDF, with duplicate removal to enhance semantic distinctiveness. Bias was measured via WEAT (Word Embedding Association Test) scores, and statistical significance was assessed through permutation testing.

Results showed that `Documentary` and `Fantasy` aligned with art films, while `Animation`, `Family`, and `Musical` were associated with commercial films—consistent with common social perceptions. PCA visualizations further illustrated how genre-level semantic structures vary across embedding models.

By systematically analyzing bias in movie-based word embeddings, this study offers insights for identifying and mitigating bias in NLP models and contributes to the development of socially responsible language technologies.

1 Introduction

These results highlight that embedding architecture significantly influences the way bias is manifested and that cultural content such as movie synopses can serve as a valuable domain for studying semantic bias in language models. In the following sections, we review prior work and present our methodology for analyzing bias in this context.

1.1 Background

Word embeddings are distributed representations of words in continuous vector spaces that capture semantic and syntactic relationships based on word co-occurrence patterns in large text corpora. Popular methods such as Word2Vec Mikolov et al. [2013], GloVe Pennington et al. [2014], and FastText Bojanowski et al. [2017] have enabled a variety of downstream NLP tasks by representing linguistic regularities as vector arithmetic.

However, because these models learn from raw textual data, they also tend to internalize and reproduce human-like biases. Such biases are often subtle and difficult to detect, yet they can significantly affect real-world applications such as search engines, translation systems, and recommendation algorithms.

1.2 Related Work

Caliskan et al. Caliskan et al. [2017] introduced the Word Embedding Association Test (WEAT), demonstrating that embeddings trained on web corpora replicate implicit human biases measured by

NeurIPS 2021 AI for Science Workshop.

psychological tools such as the Implicit Association Test (IAT). Their work showed that word vectors reflect cultural stereotypes about gender, race, and occupations.

Brunet et al. Brunet et al. [2019] further analyzed how individual documents in a training corpus contribute to the overall bias in embeddings by proposing a method to approximate the effect of corpus perturbation on bias scores. While these studies focused on general-purpose corpora like Wikipedia and news data, little attention has been given to narrative or genre-based texts such as movie synopses.

Our work extends this line of research by applying WEAT to a corpus of movie synopses and comparing the effects of different embedding models (CBoW, Skip-gram, FastText) on bias expression. In addition, we visualize the structural bias differences across models using PCA, offering new insights into how bias is encoded differently depending on the embedding architecture.

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions regarding figures, tables, acknowledgments, and references.

2 method

This study conducted word embedding-based bias analysis using a movie synopsis corpus. The overall analysis consists of five stages from morpheme extraction to embedding model comparison and bias visualization.

2.1 Morphological Analysis for Noun Extraction

First, we decomposed the synopsis text into morpheme units using the MeCab morphological analyzer. Among these, only tokens with noun parts of speech (NNG, NNP, NNB, etc.) were extracted and used as input data for word embedding learning.

2.2 Word Embedding Model Construction

Based on the corpus composed only of nouns, we learned three word embedding models: CBoW, Skip-gram, FastText. - Word2Vec-based CBoW and Skip-gram models were implemented through the ‘gensim‘ library, with embedding dimensions set to 100, window size to 5, and minimum word frequency to 3. - The FastText model aimed to improve representations.

2.3 Target and Attribute Word Set Construction

For bias analysis, we constructed two types of word sets: **target word sets** based on movie categories (art/general) and **attribute word sets** based on movie genres (e.g., drama, horror, action, etc.).

To extract representative words, we applied TF-IDF to the synopses of each category. For the **target word sets**, we selected the top 100 TF-IDF words from art and general movie synopses respectively, removed overlapping words between the two categories, and finalized 15 semantically distinctive words for each group.

For the **attribute word sets**, we also extracted the top 100 TF-IDF words from each genre's synopses. However, due to frequent overlaps between genres, we chose not to remove duplicates in this case, instead selecting the top 15 TF-IDF words per genre to better preserve the unique semantic characteristics of each genre.

2.4 WEAT Score Calculation and Visualization

To quantitatively analyze the bias inherent in word embedding spaces, we used the Word Embedding Association Test (WEAT) Caliskan et al. [2017]. WEAT measures the difference in association between two target word sets (X, Y) and two attribute word sets (A, B), providing a numerical evaluation of the presence of social bias.

The WEAT score is calculated as follows:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{1}$$

where

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}) \tag{2}$$

Here, $\cos(\vec{w}, \vec{a})$ denotes the cosine similarity between word vectors $\vec{w}$ and $\vec{a}$. Statistical significance was assessed using effect size (Cohen's d) and p-values obtained from permutation testing.

In this study, we calculated WEAT scores between movie types (target: art film word set X, commercial film word set Y) and genre-specific word sets (attribute: word sets for each genre A, B). To this end, we first extracted embedding vectors for each word from the trained CBoW model, then computed WEAT scores based on cosine similarity. We constructed (A, B) for all genre combinations and performed repeated calculations with (X, Y) fixed.

The WEAT scores were calculated using the standard formula $score(X, Y, A, B)$ and stored in a symmetric matrix format. These results were visualized as heatmaps, with colors adjusted based on the sign and magnitude of the scores to intuitively convey the direction and strength of semantic bias between genres.

For a comprehensive view of genre-wise semantic bias patterns, the full heatmaps are presented in Appendix (Figures 4–6-5).

Additionally, we sorted all genre pair WEAT scores by absolute value and extracted the top 10 most polarized pairs. This helped identify genre combinations with particularly strong bias signals, offering insights into potential social and cultural biases embedded in the data.

2.5 Embedding Comparison through PCA

We averaged the embedding vectors of representative words extracted for each genre (attribute) and reduced them to two dimensions using Principal Component Analysis (PCA).

The analysis targets were three embedding models (CBoW, Skip-gram, FastText) trained on the same word set, through which we compared the structural characteristics of the semantic spaces formed by each model.

In particular, by visually analyzing the relative distances between genres, clustering patterns, and degree of dispersion for each model, we aimed to confirm **how differences in embedding model structure affect the representation of semantic bias**.

3 Results and discussion

3.1 Movie Bias Analysis Results through WEAT Scores

Table 1 presents the top 10 genre pairs ranked by the absolute value of the WEAT scores computed using the CBOW model. The WEAT score quantitatively measures the degree of semantic bias between two genres with respect to the art film word set and the general film word set, where higher absolute scores indicate stronger bias. Statistical significance (p-value) and effect size were obtained through permutation testing.

Table 1: Top 10 genre pairs by absolute WEAT score (CBOW model) with statistical test results

Genre A	Genre B	WEAT score	Abs. Score	p-value	Effect Size
Adventure	Animation	1.1287	1.1287	0.0100	1.13
Animation	Fantasy	-1.1234	1.1234	0.0380	-1.12
Documentary	Fantasy	-1.1135	1.1135	0.0320	-1.11
Drama	etc	1.0784	1.0784	0.4960	1.08
Animation	Comedy	-1.0572	1.0572	0.0220	-1.06
Musical	Adventure	1.0549	1.0549	0.0200	-1.05
Animation	Drama	-1.0275	1.0275	0.1140	-1.03
etc	Comedy	1.0234	1.0234	0.2180	-1.02
Comedy	Musical	1.0225	1.0225	0.0480	1.02
Family	Animation	-1.0063	1.0063	0.0300	1.01

Notably, Adventure vs. Animation (WEAT = 1.1287, p = 0.0100, effect size = 1.13) and Animation vs. Fantasy (WEAT = -1.1234, p = 0.0380, effect size = -1.12) exhibited the strongest and statistically significant biases. These results suggest that Animation is more semantically associated with general films, while genres such as Fantasy and Documentary tend to align more closely with art films.

Among the top genre pairs, Animation appeared most frequently (4 out of 10), indicating its central role in the semantic distinction between art and general films. In contrast, some pairs (e.g., Drama vs. etc) showed high bias scores but lacked statistical significance (p = 0.4960), suggesting that not all observed biases are reliable.

Overall, these findings demonstrate that word embeddings capture latent genre-related biases that reflect commonly perceived distinctions between art and general films.

3.2 Structural Comparison between Embedding Models: PCA Visualization

The genre-wise average embedding vectors visualized through PCA showed distinct distributional differences across models.

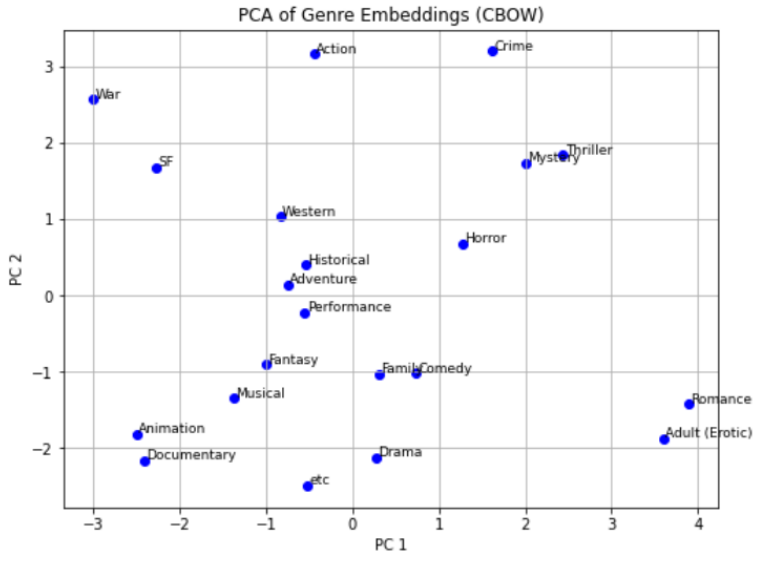

Figure 1: PCA visualization of genre embeddings using the CBoW model.

Interpretation of CBoW: In the case of the **CBoW model**, all genres were widely distributed, with both art film and commercial film-related genres existing at clearly distinguishable positions within the embedding space. This demonstrates that CBoW strongly reflects semantic differences between genres through its averaging-based learning structure and clearly learns the distinction between art and commercial films.

Furthermore, the PCA visualization in Figure 1 shows partial alignment with the WEAT analysis results in Table 1. For instance, genres such as `Animation`, which showed statistically significant association with commercial films in the WEAT test, appear relatively close to commercial film-related clusters in the embedding space. Conversely, `Documentary` and `Fantasy`, which were more strongly associated with art films in the WEAT results, are located in directions roughly consistent with art film clusters.

However, PCA also revealed genre relationships that diverged from the WEAT-based interpretation. For example, `Animation` and `Documentary`, despite having opposite bias directions in WEAT, appear closely positioned in the embedding space. Likewise, clusters such as `Family`–`Comedy` and `Fantasy`–`Musical` show strong spatial proximity, even though their bias directions differ in the WEAT results. This discrepancy illustrates that while WEAT captures directional semantic bias relative to target sets, PCA reflects overall spatial structure. Thus, both methods offer complementary insights into genre-level bias representations in CBoW embeddings.

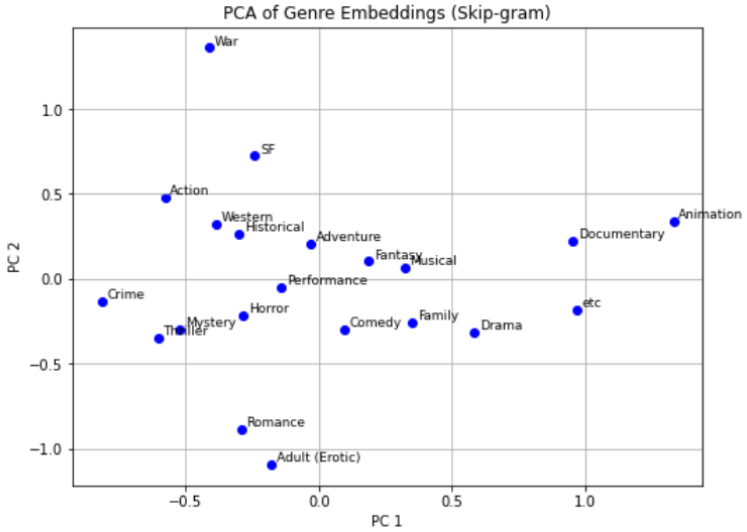

Figure 2: PCA visualization of genre embeddings using the Skip-gram model. Most genres cluster closely with low variance, indicating conservative structure.

Interpretation of Skip-gram: In contrast, the **Skip-gram model** exhibited low variance in the embedding vectors, with most genres—except for `War`, `Romance`, and `Adult (Erotic)`—clustered compactly along the PC2 axis. The PCA projection range was narrow (approximately -1 to +1), indicating smaller semantic differentiation compared to the CBoW (-4 to +4) and FastText (-3 to +3) models. This conservative structure suggests that Skip-gram embeddings tend to suppress fine-grained bias signals and emphasize co-occurrence generality over separation, which may limit its utility in revealing subtle genre distinctions.

Interpretation of FastText: The **FastText model** demonstrated the widest spatial range and highest variance among the three, reflecting its sensitivity to subword information. Genres such as `Romance` and `Adult` were positioned very closely but remained isolated from the core cluster, while semantically similar genres such as `Historical`–`Western`–`Performance`, `Comedy`–`Family`, and `Thriller`–`Mystery` formed well-defined groupings. This suggests that FastText effectively captures nuanced lexical relationships and semantic cohesion among related genres. Its flexible

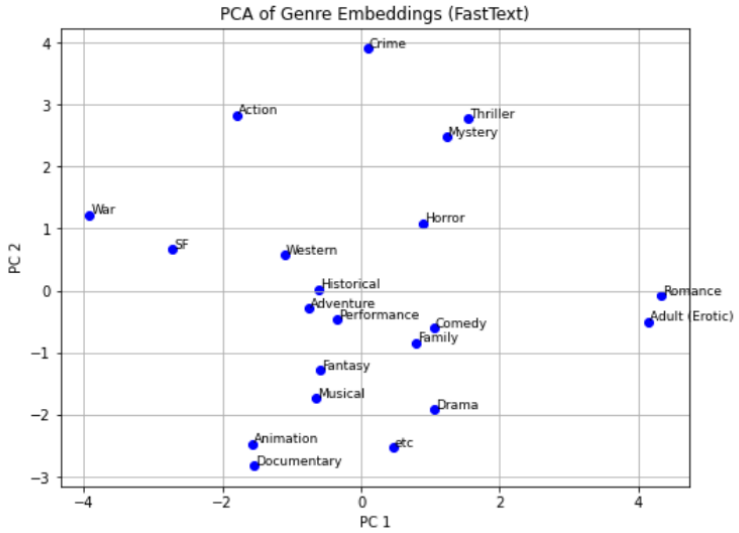

Figure 3: PCA visualization of genre embeddings using the FastText model. Balanced distribution is observed with meaningful outliers and subword sensitivity.

spatial distribution also implies that FastText embeddings may offer a richer structural representation of genre semantics, although its alignment with explicit bias signals was not evaluated in this study.

These results suggest that structural differences in word embedding models significantly influence how genre-level semantics and biases are encoded. Such differences should be carefully considered when selecting embedding models, especially in applications involving fairness or bias-sensitive tasks.

3.3 Discussion and Conclusion

This study systematically explored the semantic distinction between art and commercial films and the inter-genre bias structure by comparing various word embedding models (CBoW, Skip-gram, FastText) trained on a movie synopsis corpus.

The WEAT score analysis revealed that genres such as `Fantasy` and `Documentary` exhibited high semantic similarity with words associated with art films, while genres like `Animation`, `Family`, and `Musical` showed stronger associations with commercial film-related vocabulary. These findings align, to some extent, with commonly held public perceptions. For example, `Documentary` is often considered a genre characterized by thematic depth and artistic experimentation, whereas `Animation`, `Family`, and `Musical` are generally viewed as entertainment-oriented and more closely aligned with commercial cinema.

The initial construction of word sets was based on TF-IDF. However, the presence of overlapping words between the art and commercial film groups introduced ambiguity in the interpretation of WEAT scores. To resolve this, we extracted the top 100 TF-IDF words for each category and removed overlapping terms to derive 15 semantically distinct representative words per group.

For genre-specific attribute word sets, substantial overlap across genres was observed. Removing duplicates in this case led to sparsely populated word sets—some genres such as `Drama` contained fewer than seven usable words. Therefore, duplicates were retained to ensure semantic richness and to support more stable and reliable WEAT comparisons.

We also visualized the semantic structure of each embedding model using PCA. The CBoW model exhibited the highest degree of variance among genres, reflecting its tendency to emphasize inter-genre semantic differences. However, the PCA results did not reveal clearly separated clusters for art and commercial film genres. This discrepancy can be attributed to the differing nature of the two

methods: while WEAT quantifies directional semantic bias with respect to predefined target sets, PCA provides a global view of relative spatial positions in the embedding space.

The Skip-gram model showed the smallest variance, with most genres clustered near the center, indicating a conservative structure and limited capacity to capture meaningful genre-level distinctions. In contrast, the FastText model exhibited a moderate level of variance and effectively grouped semantically similar genres, such as `Romance–Adult (Erotic)`, `Thriller–Mystery`, and `Comedy–Family`, suggesting its strength in capturing fine-grained semantic relationships through subword-level representation. These results suggest that the architectural characteristics of word embedding models significantly influence their ability to capture cultural or genre-based semantic distinctions.

Given the limited number of documents available per genre, this study employed a TF-IDF-based approach for selecting representative words. In future work, access to larger corpora would allow for the use of more sophisticated techniques, such as Latent Semantic Analysis (LSA), to uncover deeper latent topics and construct more robust and generalizable word sets.

In conclusion, this study empirically demonstrates that both the construction of word sets and the choice of embedding model are critical factors in detecting semantic bias. Future research could extend this work by incorporating sentence-level embeddings, leveraging contextualized models such as BERT, and exploring generalization across broader domains.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2021/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to autmoatically hide this section in the anonymized submission.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR, 2019.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Appendix

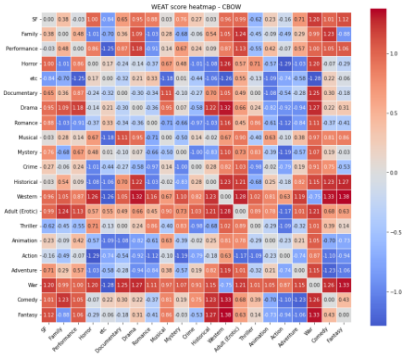

Figure 4: Heatmap of WEAT scores between genre pairs. Warmer colors indicate higher positive bias toward art films; cooler colors indicate higher bias toward commercial films.

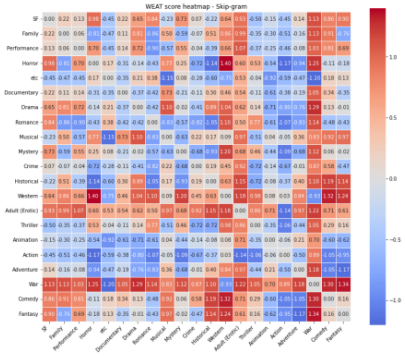

Figure 5: Heatmap of WEAT scores between genre pairs. Warmer colors indicate higher positive bias toward art films; cooler colors indicate higher bias toward commercial films.

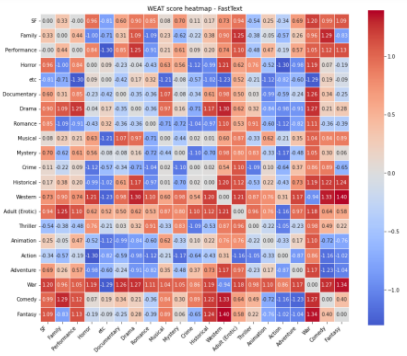

Figure 6: Heatmap of WEAT scores between genre pairs. Warmer colors indicate higher positive bias toward art films; cooler colors indicate higher bias toward commercial films.