Genre-Level Bias Analysis of Korean Movie Synopses: A Comparative Study of TF-IDF and LSA Keywords

Seongho Cho[*]
School of Computing, KAIST
mensa1229@kaist.ac.kr

Abstract

We investigate how keyword extraction methods affect semantic bias measurements between art and commercial films using the Word Embedding Association Test (WEAT). Using Korean movie synopses, we compare three methods—standard TF-IDF, frequency-filtered TF-IDF, and Latent Semantic Analysis (LSA)—to extract genre-specific keywords. Results show that each method reveals different bias patterns, with trade-offs between redundancy, representativeness, and semantic clarity. We discuss how these choices impact interpretation and suggest directions for hybrid approaches.

1 Introduction

Genre classification in film is not merely a matter of categorization—it reflects implicit cultural assumptions about what is considered artistic or commercial. Despite this, few studies have attempted to quantitatively analyze such biases, especially from the perspective of language used in film descriptions.

Existing approaches to genre analysis primarily rely on metadata, user ratings, or narrative structure. In contrast, bias detection techniques from natural language processing, such as the Word Embedding Association Test (WEAT), provide a means to uncover latent semantic biases embedded in text.

In this work, we apply WEAT to Korean movie synopses to measure the bias of genre-specific keywords toward either art or commercial films. We compare three keyword extraction methods—standard TF-IDF, frequency-filtered TF-IDF, and Latent Semantic Analysis (LSA)—and analyze how each influences the resulting bias scores across 21 genres.

Our findings demonstrate how keyword selection methods affect bias interpretation and highlight challenges in balancing representational clarity with semantic uniqueness. This study contributes to the broader understanding of linguistic bias in cultural data and provides insight for future work in interpretable media analysis.

2 Background

Word Embedding Association Test (WEAT) is a statistical method introduced by Caliskan et al. [1] to quantify bias in word embeddings. It compares the relative similarity of two sets of target words to two sets of attribute words using cosine similarity, producing a standardized effect size as the WEAT score.

[*] Work done as part of Modu Lab research 13th.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

TF-IDF (Term Frequency-Inverse Document Frequency) is a classical method for keyword extraction based on the importance of a word within a document relative to its frequency across documents [3]. While intuitive and interpretable, TF-IDF often selects high-frequency words that lack semantic specificity.

Latent Semantic Analysis (LSA) applies truncated Singular Value Decomposition (SVD) to a document-term matrix [2], capturing latent semantic structures and reducing noise. It is useful for identifying conceptually related words but may blur genre-specific boundaries.

Previous studies on genre analysis have largely focused on classification or recommendation tasks using metadata or user preferences. In contrast, our work draws from NLP bias measurement techniques to analyze genre-level semantic tendencies and their alignment with cultural perceptions of artistic or commercial value.

3 Method

In this section, we describe the dataset, preprocessing procedures, keyword extraction methods, and bias measurement technique used in our experiment. The goal of the experiment is to assess how different keyword extraction strategies affect bias measurements when applying the Word Embedding Association Test (WEAT) to film genre synopses. Through this, we aim to identify keyword selection strategies that best reflect human-perceived bias between artistic and commercial films.

3.1 dataset and preprocessing

We collected a dataset of Korean movie synopses from films produced between 2001 and August 2019. The dataset includes metadata such as synopsis, genre, and categorization into either *art films* or *commercial films*, based on classification provided by the Korea Box Office Information System (KOBIS).

The genres are divided into 21 predefined categories: *['Science Fiction', 'Family', 'Performance', 'Horror', 'Others', 'Documentary', 'Drama', 'Romance', 'Musical', 'Mystery', 'Crime', 'Historical', 'Western', 'Erotic', 'Thriller', 'Animation', 'Action', 'Adventure', 'War', 'Comedy', 'Fantasy']*. These genres were used to group and analyze the films' content, with each genre treated as a separate topic during keyword extraction.

For text preprocessing, we applied the Okt morphological analyzer from the KoNLPy library to extract only nouns from each synopsis, under the assumption that nouns carry the majority of semantic content relevant for topic and bias analysis.

To construct the word embedding space required for WEAT, we trained a Word2Vec model using the `gensim` library on the preprocessed corpus. This allowed for consistent vector representations of nouns across all genres.

3.2 keyword extraction methods

To evaluate how keyword selection strategies affect bias measurement, we extracted representative keywords from two types of corpora: (1) art vs. commercial films, and (2) 21 genre-specific film groups. For fair comparison, the number of keywords was fixed to 15 per group across all settings.

For art and commercial films, we applied **duplicate-filtered TF-IDF** as a fixed method. In contrast to genre-based extraction, which covers a wide range of categories, the keyword sets extracted from the art and commercial film corpora showed little variation across different methods. Therefore, we used a consistent keyword selection method for these two groups.

For each genre-specific group, we applied the following three keyword extraction methods:

1) Standard TF-IDF. We calculated the Term Frequency-Inverse Document Frequency (TF-IDF) score of each word w in genre document d using the formula:

$$\text{TF-IDF}(w, d) = \text{TF}(w, d) \cdot \log\left(\frac{N}{\text{DF}(w)}\right)$$

where $\text{TF}(w, d)$ is the term frequency of word w in document d, N is the total number of documents (i.e., genres), and $\text{DF}(w)$ is the number of documents in which w appears. We selected the top 15 words with the highest TF-IDF scores for each genre.

2) TF-IDF with frequency filtering. After computing TF-IDF scores as above, we filtered out any words that appeared more than four times across all genre keyword sets. This constraint was introduced to encourage genre-specific uniqueness and reduce cross-topic bias from overly generic terms.

The threshold of four was chosen empirically. During preliminary experiments, we compared multiple filtering strategies such as removing words duplicated across any genres, or appearing in more than two or three genres. We found that stricter filtering (e.g., removal thresholds below four) often led to insufficient keywords per genre or included rare terms with low frequency and weak representativeness. Setting the threshold to four provided a balance between ensuring keyword availability and maintaining their semantic relevance.

3) Latent Semantic Analysis (LSA). To extract semantically representative keywords beyond surface-level frequency, we applied Latent Semantic Analysis (LSA) using Truncated Singular Value Decomposition (SVD). We first constructed a TF-IDF matrix $X \in \mathbb{R}^{m \times n}$, where m is the number of genre documents and n is the vocabulary size, and then decomposed it as follows:

$$X \approx U \Sigma V^\top$$

We set the number of latent topics (i.e., the reduced rank of X) to 100. For each genre document, we selected the top 5 topic indices with the highest values in the corresponding row of the matrix U, representing the most relevant topics for that genre.

To prevent excessive topic overlap across genres, we excluded any topics that appeared among the top 5 in more than 5 genres. This filtering ensured genre-specific topic selection and improved keyword diversity.

From the selected 5 topics for each genre, we extracted representative words in proportion to their topic relevance: 5 words from the most relevant topic, 4 from the second, 3 from the third, 2 from the fourth, and 1 from the fifth. This weighting scheme reduced keyword duplication by assigning more influence to dominant topics while still reflecting the contribution of secondary topics. The final keyword set for each genre thus consisted of 15 words derived from its unique topic distribution.

Each of these three methods was applied independently to extract genre-specific keywords, which were subsequently used as attribute words in the WEAT bias evaluation.

3.3 bias measurement: WEAT

To quantify the semantic bias of genre-specific keywords toward either art or commercial films, we employed the Word Embedding Association Test (WEAT). WEAT evaluates the relative association between two sets of target words and two sets of attribute words based on their cosine similarity in a shared embedding space.

Given two target word sets X (art film keywords) and Y (commercial film keywords), and two attribute word sets A and B (keywords from two genres), the WEAT score is computed as follows:

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

$$\text{WEAT}(X, Y, A, B) = \frac{\mu_X - \mu_Y}{\sigma}, \quad \mu_X = \frac{1}{|X|} \sum_{x \in X} s(x, A, B), \quad \mu_Y = \frac{1}{|Y|} \sum_{y \in Y} s(y, A, B)$$

where $\cos(w, a)$ denotes the cosine similarity between embedding vectors of w and a, and σ is the standard deviation of $s(x, A, B)$ and $s(y, A, B)$ over all $x \in X$ and $y \in Y$.

In our experiment, we fixed the target groups as the representative keywords of art films (X) and commercial films (Y). For the attribute groups A and B, we selected keywords from pairs of different genres (e.g., *drama* vs. *comedy*, *documentary* vs. *action*).

The interpretation of the WEAT score is as follows: a score closer to -1 indicates that genre A is more strongly associated with commercial films than genre B, whereas a score closer to $+1$ implies that genre A is more strongly associated with art films. A score near 0 indicates a neutral or balanced association.

This design allows us to systematically compare how genre-specific language patterns semantically align with cultural perceptions of artistic versus commercial filmmaking.

4 Results

In this section, we analyze the results of different keyword extraction strategies and their influence on bias measurement and representation quality. We begin with a qualitative evaluation of the extracted keywords for each genre under different methods, followed by quantitative bias analysis using WEAT scores.

4.1 qualitative evaluation of extracted keywords

To evaluate how well each method captured representative and discriminative words per genre, we examined the top-15 keywords extracted for each genre using the three methods: *TF-IDF*, *TF-IDF with frequency filtering*, and *LSA-based extraction*.

The complete list of top-15 keywords per genre for each extraction method is provided in Appendix A.

Using the standard TF-IDF method, we observed a high degree of keyword overlap across genres. Common terms such as ” 시작”, ” 위해”, ” 자신”, ” 그녀”, ” 사랑”, ” 사람”, and ” 남자” appeared repeatedly across many genres. When calculating the proportion of overlapping keywords to the total number of extracted keywords, we found that approximately 49.5% of all keywords were duplicates. This high redundancy rate raised concerns about the representativeness of the keywords, as many of them were generic and insufficiently descriptive of genre-specific content.

In contrast, the TF-IDF method with frequency filtering (removal of words appearing in more than four genres) significantly reduced the presence of such generic terms. Genres like horror, historical, action, comedy, and fantasy particularly benefited from this filtering, as their keywords were previously dominated by frequent but uninformative terms. After filtering, more distinct and genre-relevant keywords were selected.

However, this approach also introduced limitations. For genres like family, mystery, crime, and adventure, the original TF-IDF keywords were already appropriately representative. Filtering out frequently shared words sometimes removed genuinely relevant keywords, resulting in less meaningful replacements—often low-frequency or named entities with limited semantic value. This indicates that strict filtering, while effective in reducing overlap, may also compromise the descriptive quality of keywords in certain contexts.

The LSA-based method introduced an alternative approach by selecting keywords based on latent topic representations. We selected five topics per genre, based on their relevance scores from the SVD output, and extracted 5, 4, 3, 2, and 1 keywords respectively from each topic in descending order of relevance. Although this strategy aimed to balance topic coverage and relevance, we found that some topics still overlapped significantly across genres due to the common domain of film synopses. As a result, certain genres with subtle distinctions (e.g., musical, mystery, historical, erotic, adventure, comedy) were not effectively differentiated.

Despite this, the LSA method showed notable strengths in genres where the extracted topics aligned well with genre-specific content. In these cases, the selected keywords were more semantically coherent and informative than those from the TF-IDF methods. Furthermore, the topic-based structure helped avoid the inclusion of low-information keywords (e.g., named entities), which were often present in the frequency-filtered TF-IDF output.

The LSA method also tended to produce clearer, more interpretable keywords, with less ambiguity.

Overall, the results highlight trade-offs between redundancy reduction, representational specificity, and semantic clarity among the keyword extraction methods.

4.2 weat score analysis: tf-idf-based keywords

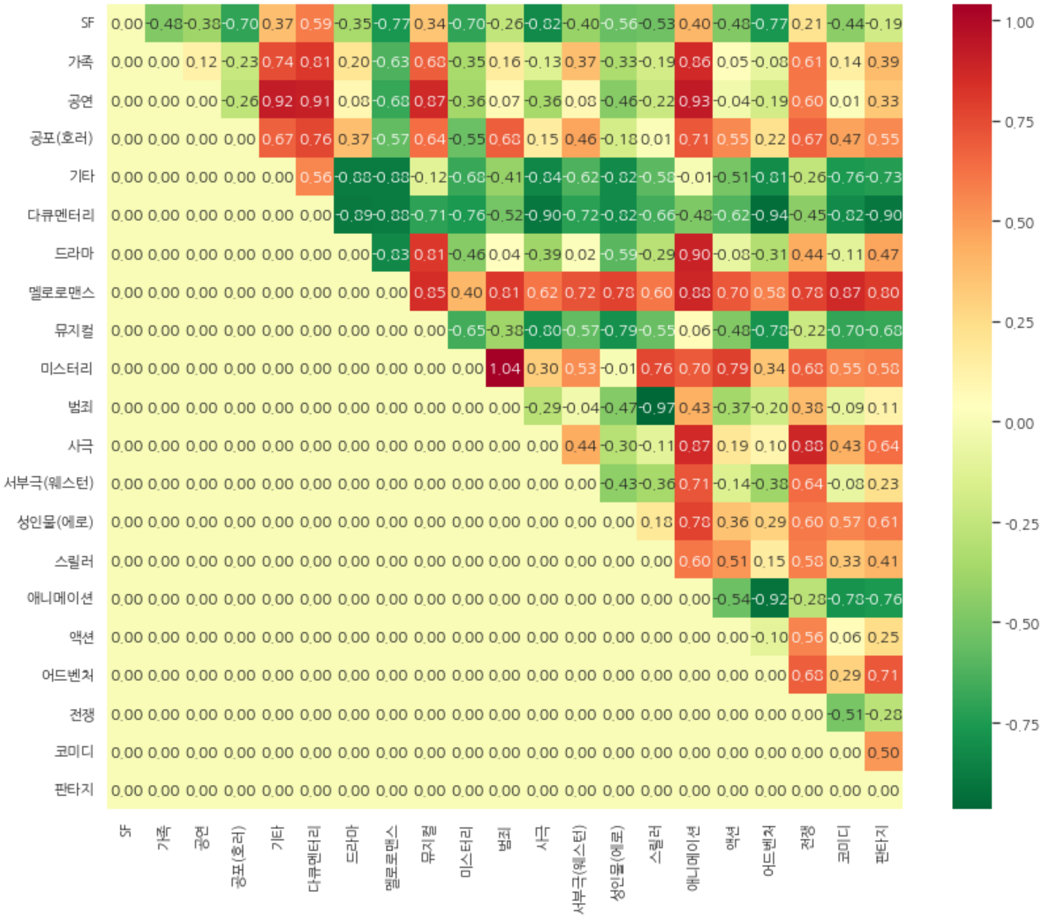

Figure 1: WEAT scores using standard TF-IDF

To analyze how the extracted keywords reflect perceived artistic or commercial tendencies, we computed WEAT scores between every possible genre pair using the standard TF-IDF keywords. The resulting heatmap is shown in Figure 1.

Several notable patterns emerge from the TF-IDF-based WEAT scores. First, *science fiction (SF)* and *drama* appear to have stronger associations with commercial films, while genres such as *horror* and *mystery* exhibit stronger associations with art films. These observations are partially consistent with prior expectations: SF is typically associated with high-budget, effects-driven commercial films, while mystery genres often emphasize narrative depth and emotional complexity, traits commonly associated with artistic cinema.

However, some results deviate from intuitive genre perceptions. For instance, drama is traditionally considered to explore human relationships and internal conflicts, which often aligns with artistic storytelling. Its strong commercial alignment in the heatmap suggests that the extracted keywords may be influenced by more mainstream or popular drama content.

Similarly, horror films are often produced as low-budget or B-grade commercial projects with strong audience appeal, making their association with art films somewhat questionable. This result may reflect the presence of more experimental or psychological horror films in the dataset.

When comparing against the *other* genre, which includes uncategorized content, several genres—including *horror*, *melodrama*, *mystery*, *crime*, and *animation*—show a consistent bias toward art films. While some of these associations are interpretable, others are less intuitive. For example, animation is frequently targeted toward children or general audiences and thus often considered commercial. Its strong leaning toward artistic films in comparison with the ”other” genre raises questions about representational skew in keyword selection.

Finally, it is surprising that the *documentary* genre—commonly regarded as closely aligned with artistic intent—does not strongly associate with art films in this setting. Although the result trends slightly toward the artistic side, its relative neutrality in the TF-IDF heatmap suggests that the keywords may not adequately capture the distinctiveness of the documentary genre.

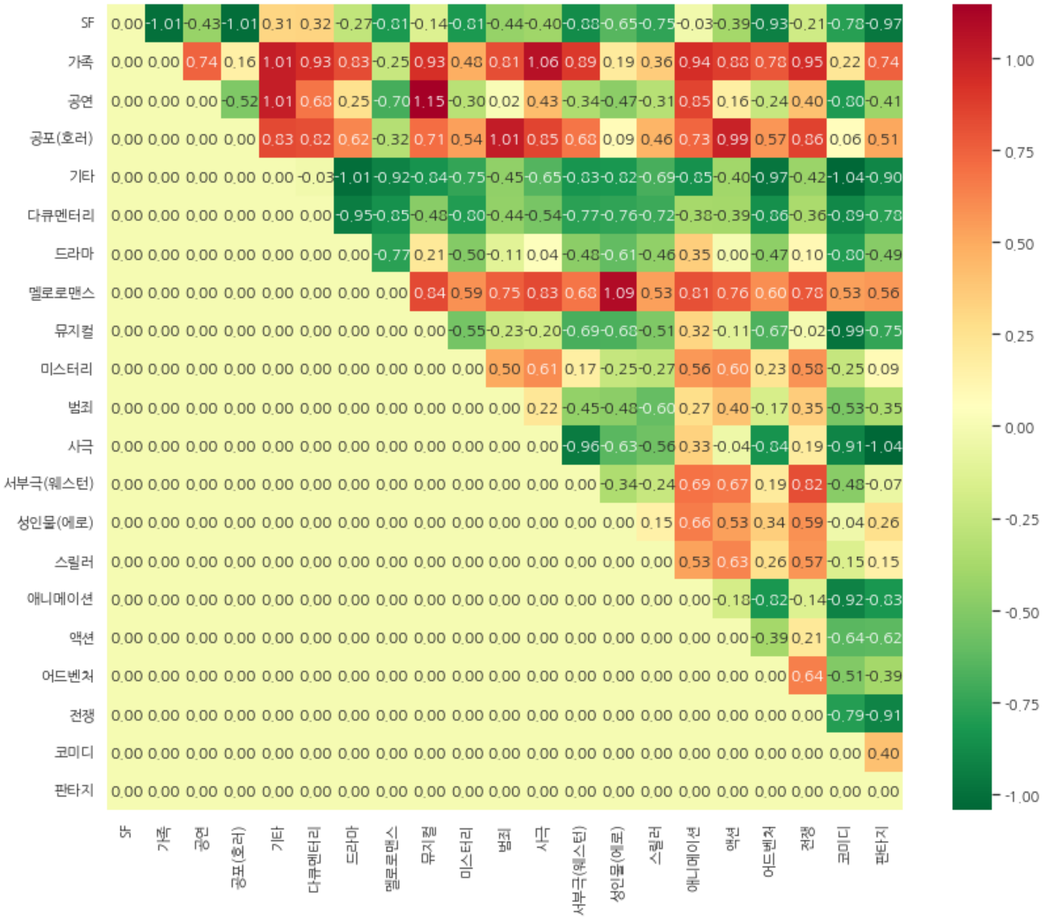

Figure 2: WEAT scores using frequency-filtered TF-IDF

4.3 weat score analysis: frequency-filtered tf-idf

The WEAT scores computed using frequency-filtered TF-IDF keywords (Figure 2) show a distribution that is broadly similar to that observed with the standard TF-IDF method. However, several key differences are worth noting.

Genres such as *family*, *fantasy*, and *comedy* display a stronger association with art films than in the previous setting. This result is somewhat unexpected, as these genres are typically aligned with commercial film production and general audience targeting. Their increased association with artistic films may reflect the impact of filtering out high-frequency, genre-independent keywords, which allowed more unique but potentially less representative words to dominate the keyword sets.

The *mystery* genre exhibits a reduced level of bias toward art films compared to the standard TF-IDF result. This shift is considered more realistic, as the mystery genre includes a substantial number of mainstream and commercial productions. The decrease in bias aligns better with its dual artistic and commercial characteristics.

Overall, the filtered TF-IDF method produced higher WEAT score magnitudes, indicating stronger bias signals. This increase is likely due to the greater differentiation in keyword sets across genres. By removing shared terms, each genre's keywords became more distinct, potentially leading to clearer association patterns with either art or commercial film categories.

These results suggest that frequency filtering can amplify detectable semantic biases, but may also introduce distortions by excluding genuinely representative but common terms.

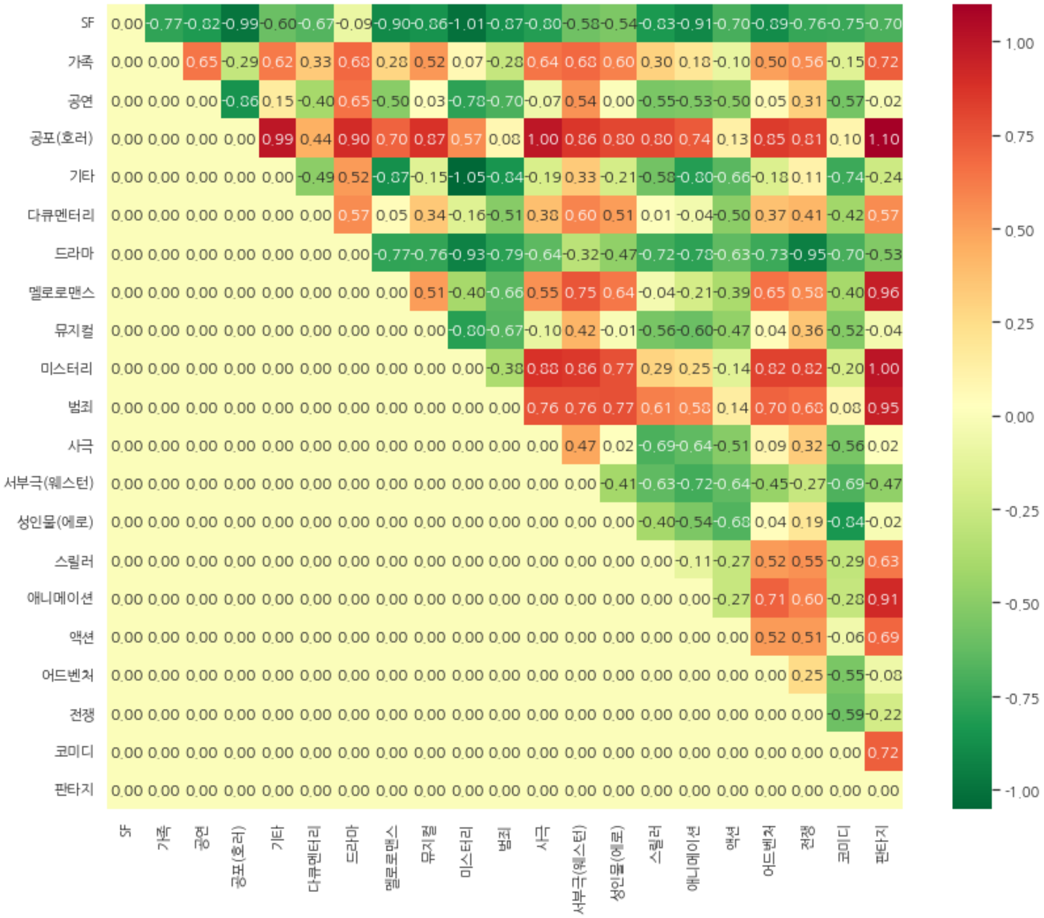

Figure 3: WEAT scores using LSA-based keywords

4.4 weat score analysis: lsa-based keywords

Figure 3 presents the WEAT scores computed using LSA-based keyword extraction. Compared to the TF-IDF-based results, the LSA method reveals similar overall trends but introduces more nuanced shifts in genre associations.

As in previous methods, *science fiction (SF)* and *drama* continue to show a strong alignment with commercial films, while genres such as *horror*, *mystery*, and *crime* lean toward artistic associations. The clear commercial alignment of SF and the artistic tendency of mystery are generally consistent with public perception, suggesting that the LSA approach maintains validity for these genres. However, drama—despite often exploring introspective and relational themes—still trends toward commercial cinema, which may reflect limitations in topic modeling to capture abstract or emotional nuance.

A particularly important shift is observed in the *documentary* genre. While TF-IDF-based scores positioned it closer to commercial films (an unexpected result), the LSA-based score shifts toward a more neutral stance. Although not strongly aligned with art films, this shift reduces the previous inconsistency and reflects improved semantic capture, even if the result still falls short of expectations.

A genre-by-genre comparison between TF-IDF and LSA methods yields additional insights:

- *Others*: Previously skewed toward commercial, now more neutral —this is a positive change, as the category lacks a clear artistic or commercial orientation. - *Melodrama*: Shifted from artistic to neutral —interpretation varies. While melodrama often appears in artistic contexts, its mainstream appeal justifies a balanced positioning. - *Crime*: Moved from neutral to artistic —this may be an unfavorable shift, as crime films span both commercial and artistic domains. - *Western*: Shifted from neutral to commercial —a plausible and favorable change, considering the genre's traditional alignment with classic commercial cinema. - *Erotic*: Transitioned from artistic to commercial —also a reasonable shift, as erotic films are more often produced for niche commercial appeal than artistic recognition. - *Animation*: Changed from commercial to artistic —this result is more ambiguous. While many artistic animations exist, the majority of animated content targets children or mass audiences, suggesting that the shift may overstate the genre's artistic tendency. - *Action*: Shifted from commercial to artistic —this is arguably a negative result. Given the high-budget, entertainment-driven nature of action films, their alignment with art films does not align well with general perceptions.

Overall, the LSA-based method yielded several improvements over TF-IDF, such as more balanced treatment of ambiguous genres and reduced noise from high-frequency generic terms. However, some semantic misalignments still persist, particularly in genres with mixed artistic and commercial elements. These results suggest that while topic modeling helps in filtering out ambiguous or generic words, it can occasionally oversimplify the genre context by focusing too rigidly on latent topics.

5 Conclusion

In this study, we investigated how different keyword extraction methods influence the measurement of semantic bias between art and commercial films using the Word Embedding Association Test (WEAT). By applying three keyword extraction strategies—standard TF-IDF, frequency-filtered TF-IDF, and LSA-based selection—we evaluated the representational quality and bias sensitivity across 21 film genres.

Our findings revealed that while standard TF-IDF suffers from excessive keyword overlap, it tends to produce intuitive results for well-separated genres such as science fiction and mystery. Frequency-filtered TF-IDF improved genre differentiation but at the cost of introducing low-meaning or named-entity terms in certain cases. LSA-based extraction offered better semantic coherence and reduced noise but occasionally misaligned with human intuition for mixed-purpose genres.

One of the primary limitations of this study lies in the trade-off between removing redundant terms and preserving semantically meaningful ones. Neither TF-IDF nor LSA alone could fully resolve this balance. Future work could explore hybrid approaches that integrate both methods, leveraging the specificity of TF-IDF with the topic sensitivity of LSA. Additionally, improvements in stopword filtering and named entity removal would likely enhance keyword quality, particularly in frequency-filtered TF-IDF scenarios.

Despite these limitations, our approach demonstrates the potential of WEAT as a tool for analyzing latent cultural bias in textual representations of genre. We hope this work contributes to the broader understanding of bias detection in cultural data and supports future research in fair and interpretable media analysis.

References

[1] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. In *Science*, volume 356, pages 183–186. American Association for the Advancement of Science, 2017.

[2] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[3] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

A Full Keyword Tables

A.1 Standard TF-IDF

A.2 Frequency-filtered TF-IDF

A.3 LSA-based Extraction

Table 1: Representative keywords per genre using standard TF-IDF

Genre	Top 15 Keywords
SF	위해, 자신, 지구, 시작, 사람, 인류, 인간, 미래, 우주, 그녀, 로봇, 세계, 모든, 박사, 우주선
Family	엄마, 아빠, 가족, 영화제, 자신, 위해, 친구, 아주르, 아버지, 시작, 그녀, 아들, 마을, 국제, 낙타
Performance	오페라, 사랑, 토스카, 실황, 올레, 자신, 카바, 그녀, 공연, 오텔로, 리골레토, 백작, 프레, 베르디, 위해
Horror	시작, 위해, 사람, 자신, 친구, 그녀, 사건, 공포, 발견, 죽음, 마을, 가족, 악령, 남자, 좀비
Others	영화제, 국제, 서울, 단편, 영화, 자신, 사람, 이야기, 그녀, 남자, 위해, 시작, 사랑, 뉴미디어, 페스티벌
Documentary	영화제, 영화, 다큐, 국제, 다큐멘터리, 사람, 이야기, 대한, 자신, 감독, 위해, 서울, 우리, 시작, 세계
Drama	자신, 영화제, 그녀, 사람, 사랑, 영화, 위해, 시작, 국제, 남자, 친구, 이야기, 엄마, 여자, 아버지
Romance	그녀, 사랑, 자신, 시작, 남편, 남자, 여자, 사람, 친구, 섹스, 위해, 마음, 결혼, 서로, 아내
Musical	뮤지컬, 사랑, 에스메랄다, 그녀, 음악, 충무로, 모차르트, 영화, 토스카, 자신, 니웨, 카바, 영화제, 바흐, 페뷔스
Mystery	사건, 그녀, 시작, 자신, 위해, 사람, 발견, 사고, 진실, 죽음, 기억, 살인, 친구, 아내, 남자
Crime	사건, 위해, 자신, 경찰, 시작, 그녀, 범죄, 조직, 살인, 사람, 마약, 형사, 남자, 모든, 살해
Historical	조선, 위해, 시작, 신기전, 사랑, 자신, 아가멤논, 황제, 그녀, 루안, 최고, 운명, 사람, 하선, 전쟁
Western	서부, 보안관, 위해, 벌린, 카우보이, 그레이프바인, 헨리, 마을, 자신, 개릿, 아이, 시작, 무법자, 프린트, 마적
Erotic	그녀, 남편, 마사지, 자신, 섹스, 관계, 영화, 정사, 남자, 위해, 시작, 여자, 유부녀, 마음, 사랑
Thriller	자신, 그녀, 사건, 시작, 위해, 사람, 살인, 남자, 발견, 아내, 경찰, 친구, 모든, 사실, 살해
Animation	애니메이션, 국제, 영화제, 친구, 인디애니페스트, 위해, 자신, 시작, 사람, 페스티벌, 서울, 이야기, 아이, 마을, 소녀
Action	위해, 자신, 시작, 조직, 사건, 사람, 그녀, 경찰, 전쟁, 모든, 목숨, 사실, 친구, 가족, 요원
Adventure	위해, 자신, 시작, 친구, 마을, 아버지, 영화, 아이, 사람, 여행, 세계, 앤트, 세상, 가족, 모험
War	전쟁, 독일군, 전투, 위해, 작전, 시작, 부대, 윈터스, 독일, 연합군, 미군, 임무, 자신, 사람, 나치
Comedy	그녀, 자신, 시작, 위해, 사랑, 사람, 친구, 영화, 남자, 여자, 영화제, 가족, 과연, 마을, 사건
Fantasy	자신, 그녀, 시작, 위해, 사람, 사랑, 요괴, 영화제, 이야기, 영화, 소녀, 남자, 인간, 세상, 마을

Table 2: Representative keywords per genre using TF-IDF with frequency filtering

Genre	Top 15 Keywords
SF	인류, 미래, 우주, 로봇, 박사, 우주선, 외계, 행성, 실험, 능력, 시스템, 생명체, 정부, 스타크, 리플리
Family	아주르, 낙타, 씨제이, 동구, 슈이트, 어머니, 마갈, 미아, 펠리칸, 벤트, 케이시, 할아버지, 엠마, 고양이, 크리스마스
Performance	오페라, 토스카, 실황, 올레, 카바, 공연, 오텔로, 리골레토, 백작, 프레, 베르디, 카르피, 비바, 왕자, 콘서트
Horror	공포, 악령, 좀비, 저주, 이후, 일행, 악몽, 병원, 파티, 유령, 귀신, 악마, 저택, 바이러스, 이사
Others	뉴미디어, 페스티벌, 독립, 아시아나, 연출, 이미지, 부산, 상영작, 지하철, 청소년, 유럽, 노인, 의도, 판타스틱, 공간
Documentary	다큐, 다큐멘터리, 한국, 환경, 사회, 노동자, 기록, 역사, 카메라, 과정, 지역, 투쟁, 인디다큐페스티발, 일상, 문제
Drama	부문, 연출, 어머니, 독립, 인생, 부산, 일상, 의도, 감정, 한국, 경쟁, 상처, 사회, 처음, 시절
Romance	섹스, 유혹, 연애, 애인, 새엄마, 불륜, 남자친구, 유부녀, 감정, 정사, 출장, 선배, 여자친구, 커플, 만난
Musical	뮤지컬, 에스메랄다, 충무로, 모차르트, 토스카, 니웨, 카바, 바흐, 페뷔스, 프롤, 모도, 카르피, 제루샤, 샤오캉, 데이비
Mystery	진실, 민혁, 미스터리, 현우, 방독면, 소설, 용의자, 공포, 여인, 추적, 의심, 사진, 조사, 이후, 랭던
Crime	범죄, 마약, 한길수, 은행, 작전, 보스, 마피아, 용의자, 추적, 프랭크, 조사, 감옥, 현장, 파푸아, 금고
Historical	조선, 신기전, 아가멤논, 황제, 루안, 하선, 윤서, 트로이, 세자, 허균, 노준, 채선, 신재효, 히파티아, 권력
Western	서부, 보안관, 벌린, 카우보이, 그레이프바인, 헨리, 개릿, 무법자, 프린트, 마적, 태구, 현상금, 분노, 버질, 랜던
Erotic	마사지, 섹스, 정사, 유부녀, 에피소드, 그린, 자위, 불륜, 욕구, 유이, 유혹, 욕구불만, 손님, 유우, 성적
Thriller	진실, 용의자, 현장, 의심, 목격, 공포, 전화, 흔적, 단서, 추적, 여인, 매력, 행동, 아파트, 조사
Animation	애니메이션, 인디애니페스트, 페스티벌, 애니, 부문, 만화, 도롱, 동물, 최강, 모험, 할아버지, 마법, 경쟁, 우주, 고양이
Action	임무, 범죄, 마약, 보스, 테러, 음모, 작전, 킬러, 인류, 무기, 암살, 부대, 갱단, 무술, 전투
Adventure	앤트, 모험, 여정, 옥자, 원주민, 보물, 동물, 윈치, 펠레, 마법, 크루소, 요정, 이름, 지역, 양말
War	독일군, 전투, 작전, 부대, 윈터스, 독일, 연합군, 미군, 임무, 나치, 병사, 이지중대, 혁리, 대원, 중위
Comedy	인생, 코미디, 섹스, 여자친구, 파티, 연애, 매력, 밴드, 문제, 준비, 삼순, 결혼식, 클럽, 만난, 대학
Fantasy	요괴, 마법, 알렉스, 순영, 판타스틱, 유령, 왕자, 공주, 남보라, 뱀파이어, 차사, 니모, 전설, 왕국, 원풍

Table 3: Representative keywords per genre using LSA

Genre	Top 15 Keywords
SF	지구, 인류, 로봇, 미래, 우주, 단편, 서울, 영화제, 뉴미디어, 부문, 자신, 연출, 사랑, 조직, 살인
Family	엄마, 아빠, 가족, 아주르, 사건, 지구, 인류, 로봇, 미래, 서부, 보안관, 벌린, 공포, 악령, 요괴
Performance	지구, 인류, 로봇, 미래, 우주, 엄마, 아빠, 가족, 아주르, 공포, 악령, 좀비, 단편, 서울, 부문
Horror	공포, 악령, 좀비, 친구, 마사지, 애니메이션, 인디애니페스트, 소녀, 요괴, 사랑, 조직, 죽음, 엄마, 아빠, 서부
Others	단편, 서울, 영화제, 뉴미디어, 남자, 공포, 악령, 좀비, 친구, 조직, 위해, 무술, 요괴, 조직, 독일군
Documentary	지구, 인류, 로봇, 미래, 우주, 독일군, 전투, 전쟁, 작전, 살인, 자신, 아내, 사랑, 조직, 공포
Drama	부문, 자신, 연출, 경쟁, 아버지, 단편, 서울, 영화제, 뉴미디어, 독일군, 전투, 전쟁, 요괴, 조직, 엄마
Romance	사랑, 조직, 죽음, 현우, 다큐, 독일군, 전투, 전쟁, 작전, 살인, 자신, 아내, 단편, 서울, 서부
Musical	애니메이션, 인디애니페스트, 소녀, 요괴, 시작, 단편, 서울, 영화제, 뉴미디어, 공포, 악령, 좀비, 엄마, 아빠, 살인
Mystery	애니메이션, 인디애니페스트, 소녀, 요괴, 시작, 엄마, 아빠, 가족, 아주르, 조직, 위해, 무술, 사랑, 조직, 단편
Crime	요괴, 조직, 사랑, 경찰, 자신, 엄마, 아빠, 가족, 아주르, 사랑, 조직, 죽음, 부문, 자신, 공포
Historical	엄마, 아빠, 가족, 아주르, 사건, 공포, 악령, 좀비, 친구, 단편, 서울, 영화제, 지구, 인류, 애니메이션
Western	서부, 보안관, 벌린, 카우보이, 그레이프바인, 독일군, 전투, 전쟁, 작전, 애니메이션, 인디애니페스트, 소녀, 지구, 인류, 단편
Erotic	독일군, 전투, 전쟁, 작전, 부대, 공포, 악령, 좀비, 친구, 서부, 보안관, 벌린, 조직, 위해, 부문
Thriller	살인, 자신, 아내, 사람, 다큐멘터리, 엄마, 아빠, 가족, 아주르, 공포, 악령, 좀비, 애니메이션, 인디애니페스트, 단편
Animation	애니메이션, 인디애니페스트, 소녀, 요괴, 시작, 살인, 자신, 아내, 사람, 사랑, 조직, 죽음, 부문, 자신, 서부
Action	조직, 위해, 무술, 임무, 뱀파이어, 요괴, 조직, 사랑, 경찰, 공포, 악령, 좀비, 살인, 자신, 사랑
Adventure	단편, 서울, 영화제, 뉴미디어, 남자, 서부, 보안관, 벌린, 카우보이, 독일군, 전투, 전쟁, 살인, 자신, 사랑
War	독일군, 전투, 전쟁, 작전, 부대, 엄마, 아빠, 가족, 아주르, 애니메이션, 인디애니페스트, 소녀, 단편, 서울, 부문
Comedy	조직, 위해, 무술, 임무, 뱀파이어, 요괴, 조직, 사랑, 경찰, 엄마, 아빠, 가족, 독일군, 전투, 서부
Fantasy	요괴, 조직, 사랑, 경찰, 자신, 애니메이션, 인디애니페스트, 소녀, 요괴, 지구, 인류, 로봇, 살인, 자신, 부문