Beats of Bias: Analyzing Lyrics with Topic Modeling and Gender Bias Measurements

Danqing Chen*, Adithi Satish*, Rasul Khanbayov*, Carolin M. Schuster, and Georg Groh

Technical University of Munich, Germany {danqing.chen, adithi.satish, rasul.khanbayov, carolin.schuster}@tum.de, grohg@in.tum.de

Abstract. This paper uses topic modeling and bias measurement techniques to analyze and determine gender bias in English song lyrics. We utilize BERTopic to cluster 537,553 English songs into distinct topics and chart their development over time. Our analysis shows the thematic shift in song lyrics over the years, from themes of romance to the increasing sexualization of women in songs. We observe large amounts of profanity and misogynistic lyrics on various topics, especially in the overall biggest cluster. Furthermore, to analyze gender bias across topics and genres, we employ the Single Category Word Embedding Association Test (SC-WEAT) to compute bias scores for the word embeddings trained on the most popular topics as well as for each genre. We find that words related to intelligence and strength tend to show a male bias across genres, as opposed to appearance and weakness words, which are more female-biased; however, a closer look also reveals differences in biases across topics.

Keywords: lyrics · gender bias · weat · embeddings · bertopic

1 Introduction

Disclaimer: Lyrics in our dataset may contain vulgar language reflected in BERTopic labels. This does not reflect the authors' opinions.

Music is integrally tied with gender identity, where lyrics, melodies, and performance styles can reflect and shape societal perceptions of gender roles, stereotypes, and experiences [2, 13, 17]. Through lyrics, artists have a way of expressing their emotions and discussing unique themes. While these themes often span a wide variety of issues, they can also propagate dangerous stereotypes and objectification [18, 21, 27, 28], pointing out the need to critically examine these gender biases that can occur in lyrics.

Natural Language Processing (NLP) techniques are well-suited for analyzing song lyrics because of their textual nature [5], facilitating unique insights into the topics and gender-related aspects. Previous research has indicated that word embeddings [4], which depict words as high-dimensional vectors, can capture linguistic biases associated with human biases [10, 26]. This inherent capability of NLP can thus be used as an advantage to explore and measure gender biases in song lyrics [6]. While there has been prior work done in identifying bias in

* These authors contributed equally to this work

lyrics [3, 5, 7], these studies mainly focus on comparing the bias that stems from artist gender, leaving a research gap in examining the gender bias implicitly present in the lyrics themselves.

Our research aims to build on this prior work by further combining topic modeling with gender bias measurements to understand how gender bias in lyrics varies thematically and across genres. Topic modeling aids in revealing the topics present in a corpus, such as song lyrics in our case [23].

In this paper, we use BERTopic [20], an advanced topic modeling technique, to take an initial look at the persistent themes in the lyrics in different genres and see how these vary across decades. Our analysis provides insights into the increasing sexualization of women in lyrics over time and the prevalence of profanity in rap music. Simultaneously, we utilize SC-WEAT analysis to quantify the gender bias in the lyrics and assess the association of various target word sets with gender-related attributes [10, 25]. Our major contributions are:

- Topic analysis, using a stratified sample, to determine cross-genre topics, recurrent themes, and charting their historical development.
- Analysis of the prevalence and variation of the gender bias in lyrics across topics and genres, using SC-WEAT scores.

2 Related Work

Song lyrics are interesting resources to determine underlying social differences, especially gender stereotypes and objectification [6, 9, 17, 28]. Prior research in NLP has shown that word embeddings pick up on cultural and gender biases inherently present in the data [8, 14, 26, 30]. In our paper, we quantify this gender bias using an extension of the Word Embedding Association Test (WEAT), the Single Category WEAT score (SC-WEAT) [5, 10, 12]. Betti et al. [5] and Boghrati and Berger [7] use the SC-WEAT score to analyze the nature of gender bias in lyrics and the differences across artist genders. However, we expand on this approach by using topic modeling to identify popular and intriguing topics. We then analyze the gender bias in the lyrics on a per-topic basis, aiming to uncover how this bias may vary across different themes.

Topic modeling clusters documents to summarize or classify them. When applied to lyrics, it provides an effective method of identifying recurring themes [16, 23]. Topic modeling algorithms like BERTopic have previously been used in gender and social science research, with Wickham [29] using the algorithm to study gender expectations on social media and their impact on suicidal ideation.

3 Experimental Setup

3.1 Data

The dataset used for the lyric analysis is a combination of artist metadata from the WASABI Song Corpus created by Fell et al. [16], and English lyrical content from Genius Song Lyrics [*]. Our lyrics dataset includes data as recent as 2022

[*] https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information

extracted from Genius, an online platform where users can upload and explain songs, poems, and even books but primarily focus on songs.

The final dataset consists of 537,553 song lyrics across five main genres and an additional miscellaneous category as described in Table 1.

Table 1: Counts of songs across genres in the dataset

Genre	**Pop**	**Rap**	**Rock**	**Country**	**R&B**	**Misc**
Counts	311,085	94,234	54,560	39,078	30,747	7,849

3.2 Topic Modeling with BERTopic

BERTopic, developed by Grootendorst [20], creates topics in four steps, which involve (i) transforming the documents into embeddings using a pre-trained language model, (ii) reducing their dimensionality, (iii) clustering and finally, (iv) deriving the topic representations from these clusters using a class-based version of TF-IDF. For the embedding process, the all-MiniLM-L6-V2 model* is used.

In BERTopic, the foundation for topic representation is built on the application of the c-TF-IDF (class-based Term Frequency-Inverse Document Frequency) algorithm, which focuses on the importance of words within topics rather than across the entire document corpus. BERTopic utilizes c-TF-IDF to weigh the word frequencies, emphasizing words that are not only frequent within a given topic but also capable of distinguishing that topic from others in the dataset.

In order to efficiently use computing resources while maintaining a good representation of the entire dataset, we use a stratified sample to train our BERTopic model by controlling for the genre. Our sample consists of 20,000 songs from each genre, as well as 7,849 "misc" entries. We then predict the topic labels for our entire lyrics corpus. The resulting topics are then used to analyze the lyrics for gender bias using SC-WEAT scores.

3.3 Bias Measurements - SC-WEAT

To analyze gender bias in lyrics, we quantify the bias by training word embeddings to compute their association scores, using an extension of the original WEAT score [10, 12], called the SC-WEAT score, which quantifies the relationship between a set of target words and two sets of attribute words [5].

SC-WEAT Score Formula: The association strength is calculated using the formula below, as proposed by Caliskan et al. [10] and used by Betti et al. [5]:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\boldsymbol{w}, \boldsymbol{a}) - \text{mean}_{b \in B} \cos(\boldsymbol{w}, \boldsymbol{b}), \tag{1}$$

$$SCWEAT(X, A, B) = \sum_{x in X} s(x, A, B), \tag{2}$$

$$d = \frac{\text{mean}_{x in X}\, s(x, A, B)}{\text{stddev}_{x in X}\, s(x, A, B)} \tag{3}$$

The cosine similarity *s(w,A,B)* is the difference between the mean cosine similarity of the word vector w to vectors in attribute sets A and B, respectively. The differential association, or effect size, is the normalized SC-WEAT score.

* https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

To compute SC-WEAT scores, we first train Word2Vec embeddings for each genre and the top topic per genre. The context-invariance of static embeddings like Word2Vec aids in analyzing average biases in the data. As our goal is to examine gender bias in the dataset rather than the model, Word2Vec, which is trained from scratch, is more suitable than contextual models like BERT [25].

We define six target sets, curated by Caliskan et al. [10] and Chaloner and Maldonado [11], and used by Betti et al. [5], in addition to two attribute sets for male and female characteristics, respectively (see Table 2). The SC-WEAT scores are calculated for each of these target sets using the aforementioned formula for each embedding model. A negative SC-WEAT score indicates a higher similarity towards the female attribute set, whereas a positive score indicates a higher similarity towards the male attribute set. The magnitude of the effect size indicates the strength of the respective bias.

Table 2: Examples of target and attribute sets used for SC-WEAT analysis. The full lists of words in these sets can be found in Betti et al. [5]

Target Set	**Examples of words in the word sets**
Pleasant	"joy", "wonderful", "love", "peace", "happy"
Unpleasant	"terrible", "hatred", "nasty", "kill", "evil"
Appearance	"thin", "gorgeous", "fat", "pretty", "beautiful", "handsome"
Intelligence	"intelligent", "genius", "smart", "brilliant", "clever"
Strength	"bold", "leader", "strong", "dominant", "power"
Weakness	"loser", "failure", "weak", "surrender", "follow"
Attribute Set	**Examples of words in the word sets**
Female	"girl", "her", "lady", "girlfriend", "mother", "she", "woman"
Male	"boy", "him", "father" "boyfriend", "dad", "he", "man"

4 Results & Discussion

4.1 Topic Analysis

The BERTopic model results in 541 topics in total, with 1.5% of the documents characterized as outliers. In Figure 1, we analyze dominant topics across music genres and their genre composition. Pop songs frequently constitute a significant portion of four distinct topics, indicating a thematic diversity in pop over other genres. A notable exception is rap, where an overwhelming 89.2% of the songs in the topic "nigga_niggas_bitch" are rap songs. While there is a prevalence of pop songs in our dataset (see Table 1), the stratification performed as detailed in Section 3.2 ensures that this does not skew the analysis.

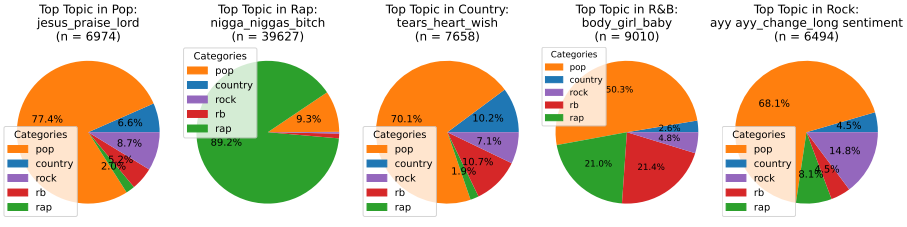


Fig. 1: Distribution of the top topic in each genre

Despite this prevalence of pop, Figure 2 reveals that the top topic in rap, "nigga_niggas_bitch", also has the highest occurrences across all genres and is relatively new, gaining popularity only from the 1990s. In essence, the top topic in pop accounts for only 1.77% of the entirety of pop songs, in contrast to rap, where the top topic accounts for 37.88% of the songs. This concentration suggests not only high popularity overall but a focused theme within rap, necessitating a deeper look into the narrative that defines a significant portion of our dataset.

Analyzing the songs in this topic, we observe the frequent occurrences of vulgar terms and profanity leading to high c-TF-IDF scores (see Figure 3a).

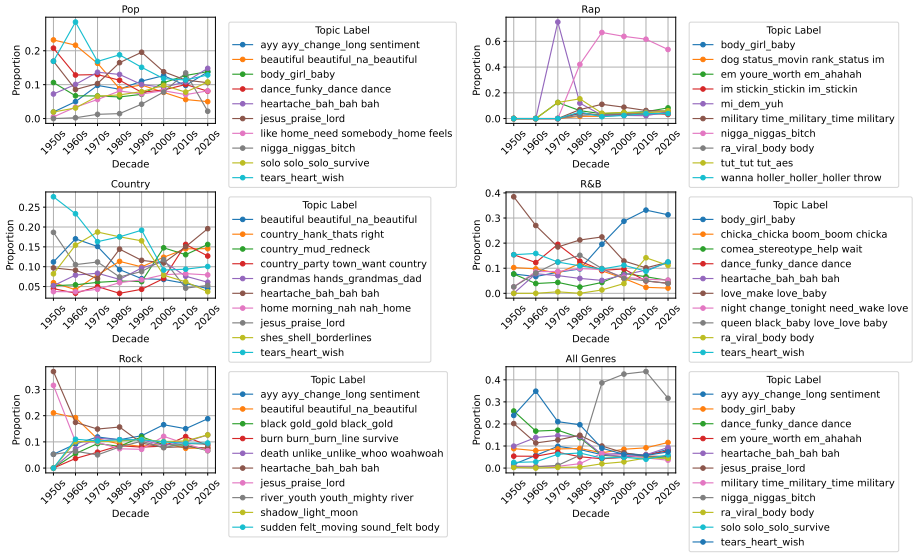


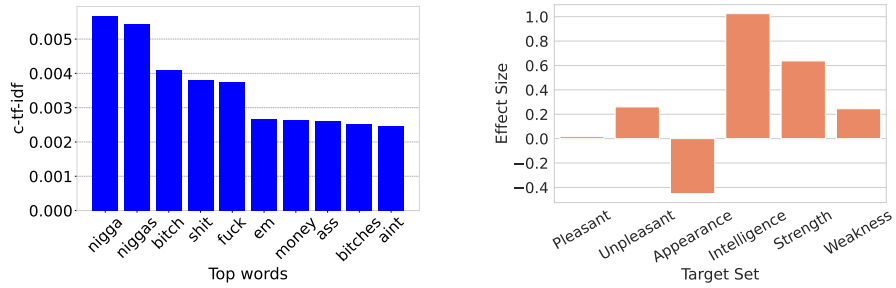
Fig. 2: Development over time of top 10 topics in each genre and overall; Decline from 2010 to 2020 can be explained by the yet still limited data for the 2020s.

Upon close examination of the lyrics of this topic exemplified by tracks like Big L's "7 Minute Freestyle" and Eminem's "Kill You," we encounter lines such as "F*ck love / All I got for hoes is hard d*ck and bubblegum." and "Slut, you think I won't choke no whore / Til the vocal cords don't work in her throat no more?!" which highlight explicit and coarse language. In alignment with this observation, Evadewi and Jufrizal [15] assert that the lyrics of rap music mostly use vulgar words, which distinguishes them from other usual English songs. When analyzing the most frequently used words in this topic and rap lyrics, we observe misogynistic lyrics that reinforce negative stereotypes or discrimination towards women, with frequent use of offensive and derogatory words such as "bitches," "sluts," and "hoes," to refer to women. In support of this statement, Adams and Fuller [1] and Grönevik [19] note that this ideology reveals itself in various forms, ranging from subtle insinuation to obvious stereotypical representations and defamations in rap. This observation about the prevalence of misogyny and profanity within rap lyrics relative to other genres also aligns with the findings published by Frisby and Behm-Morawitz [18].

Furthermore, Smiler et al. [28] also document the evolution of music content over time, shifting from themes related to romantic relationships to an increase in references to sexual behavior and objectified bodies, as evidenced in the topics in rap. This is also proven in our findings that in the top topics across successive decades, the following topics appear as trending: spanning from the 1950s to the 1960s "wonderful_sweeter years_sweeter," (due to fewer occurrences of this topic, it does not feature in Figure 2), 1960s to the 1980s "tears_heart_wish," and from 1980s to 2020s "nigga_niggas_bitch." This observation is consistent with the results reported by Hall et al. [21], who found that when comparing lyrics from 2009 to those from 1959, the occurrence of sexualized content in 2009 was over three times higher.

4.2 SC-WEAT Analysis

Employing these topics as grouping indicators, we analyze gender bias in the lyrics by calculating the SC-WEAT scores, grouped by genre, as shown in Figure 4. We observe no common trend in any genre to be male or female-biased overall; instead, they show variations in each target set.



(a) c-TF-IDF score **(b)** SC-WEAT score

Fig. 3: c-TF-IDF and SC-WEAT scores for the top topic: "nigga_niggas_bitch"

We observe that Unpleasant, Intelligence and Strength words have positive scores across genres, with higher effect sizes in rap and country. This means that the words belonging to these target sets are more closely associated with male attributes, on average, indicating a male bias in these target sets. This is in accordance with the findings in Betti et al. [5], which show the close association of Strength words with male nouns and names. Furthermore, this male bias supports previous research, which shows that men are more likely to be associated with competence ("smart," "strong," "brave," etc.) than women [6, 7].

When we take a look into the female bias, we see that average scores across genres in the Weakness target set are negative. This suggests that along with men being more closely associated with competence, women are also more likely to be associated with weakness, reinforcing traditional gender stereotypes and further increasing the gender divide in the lyrics.

In Krasse [24], the author studies the female role in pop lyrics and finds that words like "pretty," "beautiful," "ugly," "baby," etc. are very likely to precede female nouns. Our analysis supports this, with Appearance words also exhibiting

negative SC-WEAT scores in 4 out of 5 genres, suggesting that women are more closely associated with traits describing their physical appearance as opposed to their intelligence. These results also corroborate with work done in analyzing the sexualization and objectification of women in lyrics [17, 21, 22, 27].

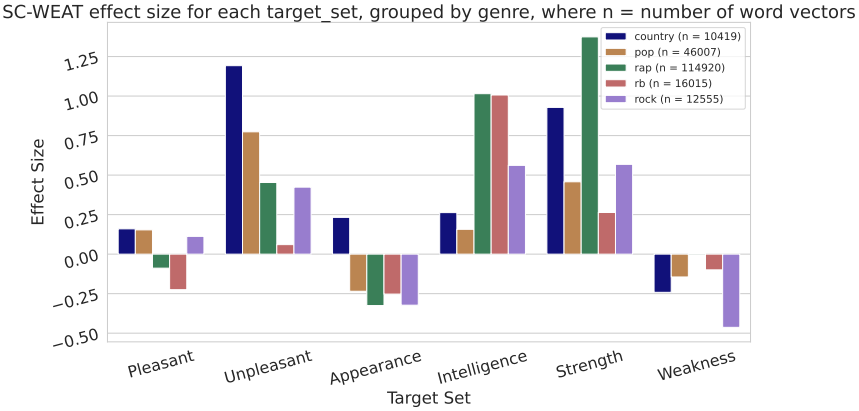


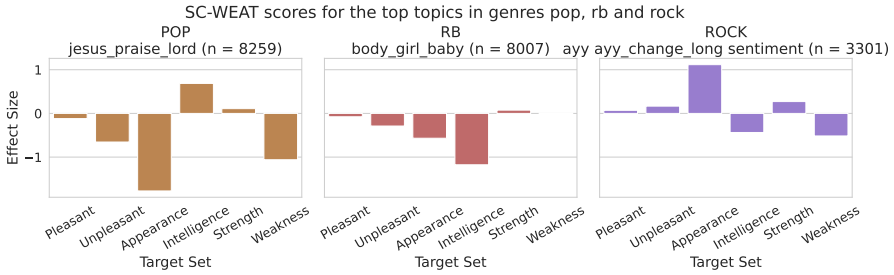
Fig. 4: The SC-WEAT effect size of the target sets in each genre. A positive score indicates male bias, whereas a negative score indicates female bias.

For a more fine-grained look, we plot the effect sizes for the top topic per genre and overall. In Figure 3b, which represents the SC-WEAT scores for the top topic overall ("nigga_niggas_bitch"), the Appearance words are female-biased, whereas Intelligence words are male-biased, highlighting the gender divide and objectification of women in this topic, as previously analyzed in Section 4.1.

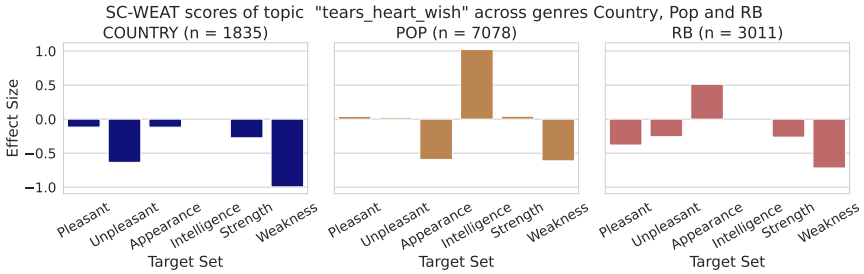
Furthermore, Figure 5a shows that the differences in association for the target sets vary across topics. Appearance words show varying levels of female bias, except in the topic "ayy ayy_change_long sentiment," where it shows a male bias, with Intelligence words showing a female bias, in contrast to the overall trend in rock. This graph underscores the need to zoom in on a per-topic level to understand how these biases can vary across topics.

Moreover, some prevalent topics in multiple genres (see Figure 1) have different biases across these genres. An example is the topic "tears_heart_wish." in country, pop, and R&B, with Figure 5b depicting the respective SC-WEAT scores. In country, the topic shows female bias regardless of the target set, with Weakness words showing the most bias. Our analysis is further supported by Rasmussen and Densley [27], who find that more than half of the country songs analyzed refer to stereotypical female gender roles and objectify women.

In Figure 5b, only Weakness words show a consistent female bias across the three genres, coinciding with our observation of women being closely associated with weakness. We also note the deviation from the average (see Figure 4) in Intelligence words for country and R&B, which show high amounts of male bias on average but negligible scores for this topic. The underlying importance of genre is underscored by the Appearance words in Figure 5b, with the set having a male bias in R&B but a female bias in pop, thereby indicating how the same topic can show different biases across different genres.



(a) Comparison of SC-WEAT bias plots of the top topics in pop, R&B, and rock.



(b) Comparison of SC-WEAT plots for "tears_heart_wish" in the country, pop, and R&B genres

Fig. 5: SC-WEAT bias plots comparison. A positive score indicates male bias, whereas a negative score indicates female bias, and n is the number of word vectors.

5 Conclusion

With the omnipresence of music in our lives and its power to evoke emotions comes the need to critically examine, identify, and understand the stereotypes and biases it can propagate and amplify. In this paper, we use BERTopic to analyze recurring themes in song lyrics and compare them across five genres (country, pop, rap, R&B, and rock) over 70 years. We analyze the gender bias prevalent in the lyrics using SC-WEAT scores to understand how the bias varies across topics and genres.

Controlling for the genre, we use a stratified sample to fit the BERTopic model to ensure adequate representation of the dataset. The top topic with the most occurrences, as labeled by the BERTopic model, is "nigga_niggas_bitch"; analysis of the lyrics reveals the prominence of misogyny and profanity. Although our dataset spans from the 1950s to the 2020s, this topic only begins to emerge in the 1990s. The trending topics before the 90s include "tears_heart_wish" and "wonderful_sweeter_years_sweeter.", revealing the shift in theme from romance to an increase in sexualization and profanity in song lyrics [21, 28].

Furthermore, the SC-WEAT results suggest the presence of an implicit gender bias in the lyrics, with words associated with Weakness and Appearance more frequently biased towards women, whereas Intelligence words exhibit a male bias. That Appearance words are largely female-biased further lends credence to our initial observation of the sexualization of women in songs [17, 21, 27]. However, when we look at the scores on a per-topic and per-genre basis, we see that different topics exhibit different kinds of bias, and sometimes, the same topic shows

variations in other genres as well, highlighting the necessity of this zoomed-in analysis. This combination of topic analysis with bias measurements provides a framework for understanding the thematic components of lyrics and how they can reinforce gender roles.

Looking towards future work, several areas emerge as critical for further exploration. Firstly, this study exclusively analyzes English-language songs. Extending this to other languages could potentially broaden the scope and applicability of the findings. Secondly, treating gender as binary overlooks the spectrum of gender identities, suggesting a need for research into gender diversity in music. Additionally, BERTopic assigns only a single topic per song, overlooking the potential presence of multiple topics; moreover, interpreting the generated topics is complex, particularly in large corpora where manual supervision is often not feasible. Finally, lyrics, with their contextual nuances and potential sarcasm, make bias identification challenging. Addressing these limitations could offer an even more nuanced and comprehensive perspective on the intersection of music, culture, and gender bias.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Adams, T. M. and Fuller, D. B. (2006). The words have changed but the ideology remains the same: Misogynistic lyrics in rap music. *Journal of Black Studies*, 36(6):938–957.

[2] Alexander, S. (1999). The gender role paradox in youth culture: An analysis of women in music videos. *Michigan Sociological Review*, pages 46–64.

[3] Anglada-Tort, M., Krause, A. E., and North, A. C. (2021). Popular music lyrics and musicians' gender over time: A computational approach. *Psychology of Music*, 49(3):426–444.

[4] Bengio Y, Ducharme R, V. P. (2000). A neural probabilistic language model. In *NIPS*, volume 13, pages 932–938. MIT Press.

[5] Betti, L., Abrate, C., and Kaltenbrunner, A. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1):10.

[6] Boghrati, R. and Berger, J. (2022). Quantifying gender bias in consumer culture. *CoRR*, abs/2201.03173.

[7] Boghrati, R. and Berger, J. (2023). Quantifying cultural change: Gender bias in music. *Journal of Experimental Psychology: General*.

[8] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

[9] Bretthauer, B., Zimmerman, T. S., and Banning, J. H. (2007). A feminist analysis of popular music: Power over, objectification of, and violence against women. *Journal of Feminist Family Therapy*, 18(4):29–51.

[10] Caliskan, A., Bryson, J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

[11] Chaloner, K. and Maldonado, A. (2019). Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32.

[12] Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., and Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.

[13] Colley, A. (2008). Young people's musical taste: Relationship with gender and gender-related traits 1. *Journal of applied social psychology*, 38(8):2039–2055.

[14] Durrheim, K., Schuld, M., Mafunda, M., and Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, 62(1):617–629.

[15] Evadewi, R. and Jufrizal, J. (2018). An analysis of english slang words used in eminem's rap music. *English Language and Literature*, 7(1).

[16] Fell, M., Cabrio, E., Tikat, M., Michel, F., Buffa, M., and Gandon, F. (2023). The wasabi song corpus and knowledge graph for music lyrics analysis. *Language Resources and Evaluation*, 57(1):89–119.

[17] Flynn, M. A., Craig, C. M., Anderson, C. N., and Holody, K. J. (2016). Objectification in popular music lyrics: An examination of gender and genre differences. *Sex roles*, 75:164–176.

[18] Frisby, C. M. and Behm-Morawitz, E. (2019). Undressing the words: Prevalence of profanity, misogyny, violence, and gender role references in popular music from 2006-2016. *Media Watch*, 10(1):5–21.

[19] Grönevik, K. (2013). The depiction of women in rap and pop lyrics.

[20] Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[21] Hall, P., West, J., and Hill, S. (2011). Sexualization in lyrics of popular music from 1959 to 2009: Implications for sexuality educators. *Sexuality Culture*, 16.

[22] Karsay, K., Matthes, J., Buchsteiner, L., and Grosser, V. (2019). Increasingly sexy? sexuality and sexual objectification in popular music videos, 1995–2016. *Psychology of popular media culture*, 8(4):346.

[23] Kleedorfer, F., Knees, P., and Pohle, T. (2008). Oh oh oh whoah! towards automatic topic detection in song lyrics. In *Ismir*, pages 287–292.

[24] Krasse, L. (2019). A corpus linguistic study of the female role in popular music lyrics.

[25] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

[26] Qin, X. and Tam, T. (2023). Stereotype content dictionary: A semantic space of 3 million words and phrases using google news word2vec embeddings. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 12–22. Springer.

[27] Rasmussen, E. E. and Densley, R. L. (2017). Girl in a country song: Gender roles and objectification of women in popular country music across 1990 to 2014. *Sex Roles*, 76:188–201.

[28] Smiler, A., Shewmaker, J., and Hearon, B. (2017). From "i want to hold your hand" to "promiscuous": Sexual stereotypes in popular music lyrics, 1960–2008. *Sexuality and Culture*, 21:1–23.

[29] Wickham, E. N. (2023). Girlbosses, the red pill, and the anomie and fatale of gender online: Analyzing posts from r/suicidewatch on reddit. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 195–212.

[30] Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. (2019). Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.