Tuning Into Bias: A Computational Study of Gender Bias in Song Lyrics

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

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

Abstract

The application of text mining methods is becoming increasingly prevalent, particularly within Humanities and Computational Social Sciences, as well as in a broader range of disciplines. This paper presents an analysis of gender bias in English song lyrics using topic modeling and bias measurement techniques. Leveraging BERTopic, we cluster a dataset of 537,553 English songs into distinct topics and analyze their temporal evolution. Our results reveal a significant thematic shift in song lyrics over time, transitioning from romantic themes to a heightened focus on the sexualization of women. Additionally, we observe a substantial prevalence of profanity and misogynistic content across various topics, with a particularly high concentration in the largest thematic cluster. To further analyse gender bias across topics and genres in a quantitative way, we employ the Single Category Word Embedding Association Test (SC-WEAT) to calculate bias scores for word embeddings trained on the most prominent topics as well as individual genres. The results indicate a consistent male bias in words associated with intelligence and strength, while appearance and weakness words show a female bias. Further analysis highlights variations in these biases across topics, illustrating the interplay between thematic content and gender stereotypes in song lyrics.

1 Introduction

Disclaimer: Lyrics in the dataset may include explicit or vulgar language, which is inherently reflected in the topic labels generated by the BERTopic model. This does not represent the views or opinions of the authors.

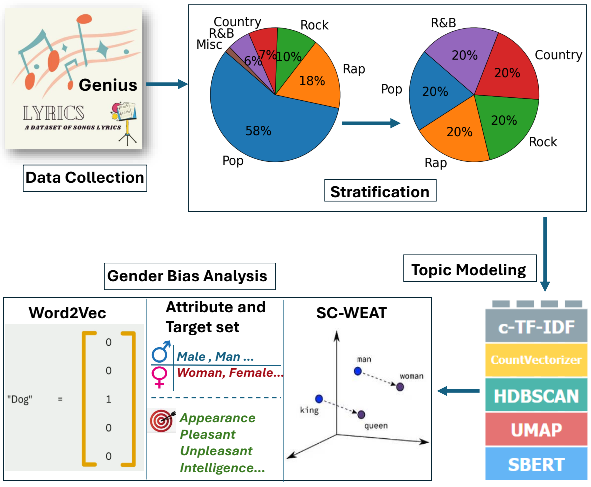


Figure 1: Detailed workflow including data collection, topic modeling, and SC-WEAT.

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 (Flynn et al., 2016; Colley, 2008; Alexander, 1999). 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 (Rasmussen and Densley, 2017; Hall et al., 2011; Frisby and Behm-Morawitz, 2019; Smiler et al., 2017a), pointing out the need to critically examine these gender biases that can occur in lyrics.

Natural Language Processing (NLP) techniques provide a robust framework for analyzing song lyrics by leveraging their underlying textual structure to extract thematic patterns and gender-associated linguistic representations (Betti et al., 2023). In particular, word embeddings (Bengio Y, 2000), which encode lexical items as dense, high-dimensional vectors within a continuous space, have been shown to effectively capture and encode latent linguistic biases that align with human cognitive associations (Caliskan et al., 2017; Qin and Tam, 2023). This representational property renders word embeddings a powerful com-

[*] These authors contributed equally to this work

Proceedings of LaTeCH-CLfL 2025, pages 117–129
May 4, 2025 ©2025 Association for Computational Linguistics

putational tool for systematically quantifying and analyzing gender biases embedded within lyrical discourse (Boghrati and Berger, 2022).

While previous research has primarily analyzed gender bias at the artist level by comparing the lyrics of songs performed by male and female artists (Anglada-Tort et al., 2021; Betti et al., 2023; Boghrati and Berger, 2023), this study does not differentiate based on the artist's gender. Instead, we focus solely on examining bias within the lyrics themselves. By integrating topic modelling with quantitative bias measurement, this approach facilitates a granular analysis of gender bias across themes and genres, utilizing NLP to bridge the gap in Humanities and Social Sciences to analyze complex text-based artefacts and their sociocultural implications.

Topic modeling is a powerful technique for uncovering the underlying themes within a corpus, such as song lyrics in our study (Kleedorfer et al., 2008). In this paper, we employ BERTopic (Grootendorst, 2022), a state-of-the-art topic modeling method, to analyze persistent lyrical themes across various genres and examine their evolution over multiple decades. This approach enables us to uncover critical insights, including the increasing sexualization of women in song lyrics over time and the notable prevalence of profanity, particularly in rap music. While the topic model provides a broad overview of the gender bias in lyrics, we also take a more fine-grained look into this bias by applying the SC-WEAT analysis to quantify it and evaluate the associations of specific target word sets with gender-related attributes (Mikolov et al., 2013; Caliskan et al., 2017). Our major contributions, as depicted in the workflow diagram in Figure 1, are:

- Conducting topic analysis on a stratified sample of song lyrics to identify cross-genre themes, recurrent topics, and the historical evolution of gender bias.
- Evaluating the prevalence and variation of gender bias in lyrics quantitatively across topics and genres through the computation of SC-WEAT scores.

2 Related Work

The intersection of music and natural language processing (NLP) has been the focus of extensive research, encompassing tasks such as mood classification, music transcription, lyrics and melody generation, among others (Laurier et al., 2008; Benetos et al., 2018; Chen and Lerch, 2020; Yu et al., 2021). Music — and, by extension, lyrics — constitutes a valuable resource for investigating underlying societal dynamics, particularly in the context of gender stereotypes and objectification (Flynn et al., 2016; Bretthauer et al., 2007; Smiler et al., 2017b; Boghrati and Berger, 2022).

Previous research has demonstrated that word embeddings are inherently susceptible to capturing and, in some cases, amplifying the social biases present in the data from which they are derived (Hovy and Prabhumoye, 2021). A well-known example provided by Bolukbasi et al. (2016) illustrates that the word embedding for “man” is more closely associated with “programmer,” while “woman” is linked to “homemaker.” Similarly, the findings of Durrheim et al. (2023) and Zhao et al. (2019) reveal that word embeddings encode implicit cultural and gender biases, even when such biases are not explicitly stated in the source data. This body of work highlights the critical importance of examining and addressing biases embedded in linguistic representations, especially when applied to cultural artifacts such as song lyrics.

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) (Caliskan et al., 2017; Charlesworth et al., 2021; Betti et al., 2023). The SC-WEAT score is also used by Betti et al. (2023) and Boghrati and Berger (2023) 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 as well as per-genre basis, aiming to uncover how this bias may vary across different themes.

Topic modeling is a widely used technique for clustering documents to summarize or classify them, enabling the identification of underlying social patterns within the data (Egger and Yu, 2022). When applied to song lyrics, it serves as an effective approach for uncovering recurring themes (Kleedorfer et al., 2008; Fell et al., 2023; Devi and Saharia, 2020; Karamouzi et al., 2024). While Latent Dirichlet Allocation (LDA) remains one of the most common methods for topic modeling, recent findings by Gan et al. (2023) demonstrate that BERTopic, introduced by Grootendorst (2022), outperforms traditional approaches by producing more distinctive and interpretable clusters.

BERTopic has also been successfully applied in gender and social science research. For example, Nakajima Wickham (2023) utilized the algorithm to examine gender expectations on social media and their influence on suicidal ideation. This demonstrates BERTopic's utility in the clustering of categories that are meaningful to societal and cultural dynamics.

3 Experimental Setup

3.1 Data

The dataset used for the lyric analysis is a combination of song metadata from the WASABI Song Corpus created by Fell et al. (2023), and English lyrical content from Genius Song Lyrics [1]. Our lyrics dataset includes data as recent as 2022 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.

Genre	Counts (% of dataset)
Pop	311,085 (58%)
Rap	94,234 (18%)
Rock	54,560 (10%)
Country	39,078 (7%)
R&B	30,747 (6%)
Misc	7,849 (1%)

Table 1: Counts of songs across genres in the dataset.

3.2 Topic Modeling with BERTopic

BERTopic leverages transformers to create clusters, providing more interpretable topic representations compared to traditional methods (Grootendorst, 2022). The algorithm 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 our analysis, we use the default configuration of BERTopic, which utilizes (i) all-Mini-LM-L6-V2 [2], (ii) UMAP, (iii) HDBSCAN and (iv) c-TF-IDF for the four steps mentioned above [3].

BERTopic leverages c-TF-IDF (class-based Term Frequency-Inverse Document Frequency) to represent topics by weighting words based on their importance within a topic rather than across the entire corpus (Grootendorst, 2022). This approach emphasizes words that are not only frequent within a given topic but also capable of distinguishing that topic from others in the dataset. To optimize computational resources while preserving dataset representativeness, we train the BERTopic model on a stratified sample comprising 20,000 songs per genre and 7,849 "misc" entries. The model then predicts topic labels for the full corpus, which are subsequently analyzed 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 from scratch to compute their association scores, using an extension of the original WEAT score (Caliskan et al., 2017; Charlesworth et al., 2021), called the SC-WEAT score, which quantifies the relationship between a set of target words and two sets of attribute words (Betti et al., 2023).

SC-WEAT Score Formula: The association strength is calculated using the formula below, as proposed by Caliskan et al. (2017) and used by Betti et al. (2023):

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

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

$$d = \frac{\text{mean}_{xinX}\, s(x, A, B)}{\text{stddev}_{xinX}\, s(x, A, B)} \quad (3)$$

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.

To compute SC-WEAT scores, we train Word2Vec embeddings for each genre and the top topic within each genre. Static embeddings, such as Word2Vec, are well-suited for analyzing aggregate biases within the data (Caliskan et al.,

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

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://maartengr.github.io/BERTopic/algorithm/algorithm.html

Target Set	**Examples of words in the word sets**
Pleasant	“joy”, “wonderful”, “love”, “peace”
Unpleasant	“terrible”, “hatred”, “nasty”, “kill”
Appearance	“thin”,“gorgeous”,“fat”,“pretty”
Intelligence	“intelligent”,“genius”,“brilliant”
Strength	“bold”,“leader”,“strong”,“power”
Weakness	“loser”,“failure”,“weak”,“follow”
Attribute Set	**Examples of words in the word sets**
Female	“girl”,“her”,“woman”,“girlfriend”
Male	“boy”,“him”,“man”,“boyfriend”

Table 2: Examples of target and attribute sets used for SC-WEAT analysis. The full lists of words, curated by Betti et al. (2023), can be found in Table 3 and Table 4 in the Appendix.

2017; Betti et al., 2023). As the objective is to examine gender bias inherent in the dataset rather than the model itself, Word2Vec—trained from scratch—is more appropriate than contextual models like BERT (Mikolov et al., 2013).

We define six target sets, curated by Caliskan et al. (2017) and Chaloner and Maldonado (2019), which are used by Betti et al. (2023), 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.

4 Results & Discussion

4.1 Topic Analysis

The BERTopic model identifies a total of 541 topics, with 1.5% of documents classified as outliers. Figure 2 illustrates the most salient topics along with their genre distributions, representing the genre composition of songs assigned to each topic label, where each label is generated based on the most representative terms, constructed using the top three words with the highest c-TF-IDF values.

While the figure shows the composition of the top topics in each genre, it reveals the dominant influence of pop in other genres as well. For instance, in addition to the top topic within pop, the top topics in country (*“tears_heart_wish”*), R&B (*“body_girl_baby”*) and rock (*“ayy ayy_change_long_sentiment”*) are also largely shaped or consist of pop songs. This indicates greater thematic diversity of pop songs, whereas rap exhibits a strong thematic concentration, with 89.2% of songs in *“nigga_niggas_bitch”* belonging to the rap genre. While pop is the most prevalent genre in the dataset (see Table 1), this imbalance is mitigated by the stratified sampling approach outlined in Section 3.2, ensuring a more balanced genre representation in the analysis.

Despite the prevalence of pop music in the dataset, Figure 3 shows that the most prominent topic in rap, *“nigga_niggas_bitch”*, has the highest frequency across all genres and emerged predominantly in the 1990s. Analyzing the distribution of top topics within each genre highlights a stark disparity: the top topic in pop accounts for only 1.77% of all pop songs, whereas in rap, the top topic represents 37.88% of the genre. This significant concentration indicates the dominant popularity and thematic specificity of this topic within rap, accounting for a substantial portion of the dataset.

This pronounced disparity emphasizes the distinctive narrative centrality of the top topic in rap compared to pop, necessitating a more detailed investigation into its linguistic and cultural characteristics. An analysis of the lyrics within this topic reveals a frequent occurrence of vulgar language and profanity, as evident from the c-TF-IDF scores (see Figure 4). These observations highlight the thematic uniqueness of rap and underline the importance of further examining the social and cultural implications embedded within its lyrical content.

A detailed qualitative analysis of the lyrics within this topic, exemplified by tracks such as *Big L’s 7 Minute Freestyle* and *Eminem’s Kill You*, reveals a prevalent use of explicit and coarse language. Notable lyrical excerpts, including *“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?!”*, exemplify this linguistic trend. These findings align with the argument presented by Evadewi and Jufrizal (2018), who contend that rap music lyrics are distinguished by the frequent incorporation of vulgar and explicit language, setting them apart from other English-language musical genres. Furthermore, a quantitative analysis of word frequency within this topic and across rap lyrics

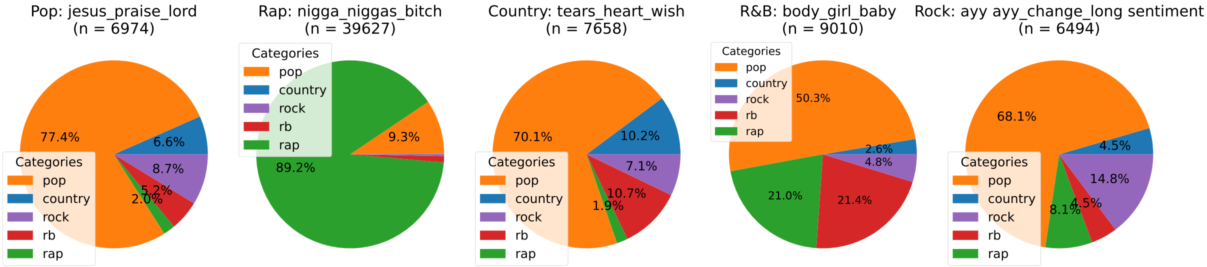


Figure 2: Distribution of the top topic in each genre, with (n) representing the number of songs associated with that topic. As shown, the top topic in each genre often includes a significant proportion of songs from other genres, indicating genre overlap in topic composition.

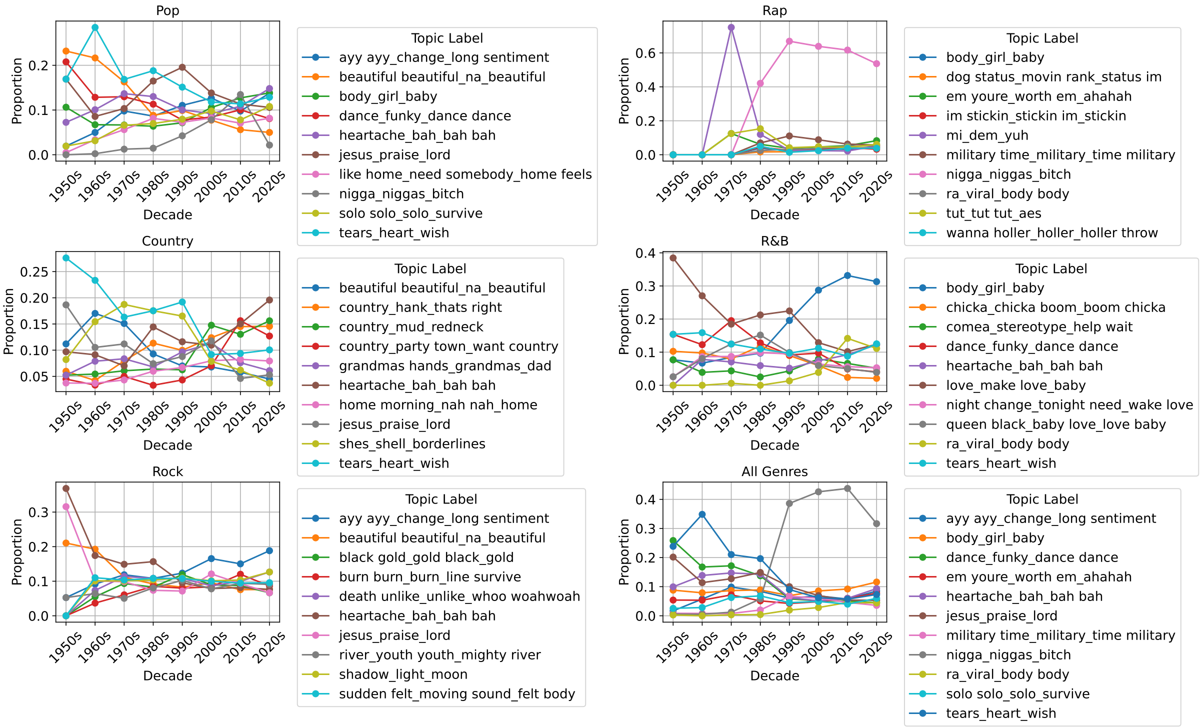


Figure 3: Development over time of top 10 topics in each genre and overall; decline from 2010 to 2020 can be explained by the yet still limited data for the 2020s.

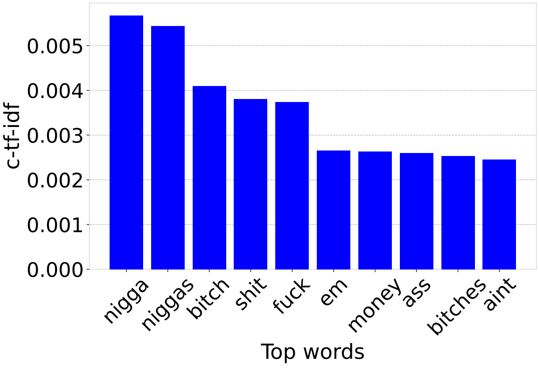


Figure 4: c-TF-IDF score for the overall top topic: *"nigga_niggas_bitch"*.

underscores the recurrent presence of misogynistic terminology, which serves to reinforce negative gender stereotypes and perpetuate discriminatory narratives. In particular, derogatory terms such as 'bitches,' 'sluts,' and 'hoes' frequently appear in reference to women, reflecting broader patterns of gendered linguistic bias within this lyrical subdomain. This observation is further corroborated by Adams and Fuller (2006) and Grönevik (2013), who highlight that such ideologies manifest through a spectrum of expressions, from subtle insinuations to obvious stereotypical representations and defamatory language within rap lyrics. Additionally, the higher prevalence of misogyny and profanity in rap lyrics, compared to other genres, aligns with the findings of Frisby and Behm-Morawitz (2019), who document similar patterns in their comprehensive analyses.

Furthermore, Smiler et al. (2017a) also documented the evolution of music content over time, shifting from themes related to romantic relationships to an increase in references to sexual behaviour 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: *"wonderful_sweeter years_sweeter"*, spanning from the 1950s to the 1960s, (due to fewer occurrences of this topic, it does not feature in Figure 3), *"tears_heart_wish"*, from 1960s to the 1980s, and *"nigga_niggas_bitch"* from 1980s to 2020s. This observation is consistent with the results reported by Hall et al. (2011), 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 5. We observe no common trend in any genre to be male or female-biased overall; instead, they show variations in each target set.

We observe that Unpleasant, Intelligence, and Strength words exhibit positive SC-WEAT scores across all genres, with notably higher effect sizes in rap and country. This indicates that these target sets are more closely associated with male attributes on average, reflecting a pronounced male bias. These findings align with prior research by Betti et al. (2023), which highlights the strong association between Strength words and male nouns or names. Furthermore, the observed male bias aligns with prior research indicating that men are more frequently associated with attributes related to competence, such as 'smart,' 'strong,' and 'brave,' in contrast to women (Boghrati and Berger, 2022, 2023).

A systematic analysis of female bias within song lyrics reveals that the Weakness target set consistently exhibits negative SC-WEAT scores across multiple genres. This trend suggests that, in parallel with the stronger association of men with competence-related attributes, women are more frequently linked to concepts of weakness. Such linguistic patterns reinforce entrenched gender stereotypes, thereby perpetuating and amplifying gendered asymmetries in lyrical discourse.

This phenomenon aligns with prior findings by Liu et al. (2023), which highlight the prevalence of gender stereotypes in media, such as the association of men with strength and women with appearance, particularly in contexts like video games. Similarly, the corpus-based study by Krasse (2019) on pop lyrics identifies a pronounced linguistic pattern wherein adjectives such as "pretty," "beautiful," "ugly," and "baby" frequently precede female nouns. Our empirical analysis substantiates these findings, revealing that Appearance-related words consistently yield negative SC-WEAT scores across four out of five musical genres. This trend highlights the predominant linguistic association of women with attributes linked to physical appearance rather than intellectual or competence-related qualities. These results are consistent with prior research documenting the pervasive sexualization and objectification of women in song lyrics (Flynn et al., 2016; Hall et al., 2011; Karsay et al., 2019; Rasmussen and Densley, 2017), further illustrating how this cultural medium serves to reinforce and perpetuate traditional gender stereotypes.

For a more granular analysis, we compute SC-WEAT scores for the top topic in each genre and overall. Figure 6 visualizes the scores for the top overall topic (*"nigga_niggas_bitch"*), where Appearance words exhibit a strong female bias, while Intelligence words show a marked male bias. These findings reinforce the gender divide and the objectification of women within this topic, as discussed in Section 4.1.

Furthermore, Figure 7 illustrates that the biases associated with target sets vary across topics. Notably, Appearance words generally exhibit a female bias; however, in the topic *"ayy ayy_change_long sentiment"*, they display a male bias, while Intelligence words show a female bias—contrasting with the overall trend observed in the rock genre (refer to Figure 5). These findings emphasize the importance of topic-specific analysis to capture the nuanced variations in biases across different topics, which might otherwise be obscured in genre-level aggregations.

Moreover, certain prevalent topics that appear across multiple genres exhibit differing biases depending on the genre (see Figure 2). For instance, the topic *"tears_heart_wish"*, which is present in the country, pop, and R&B genres, demonstrates distinct SC-WEAT scores for each genre, as shown in Figure 8. In the country genre, this topic consistently displays a female bias across

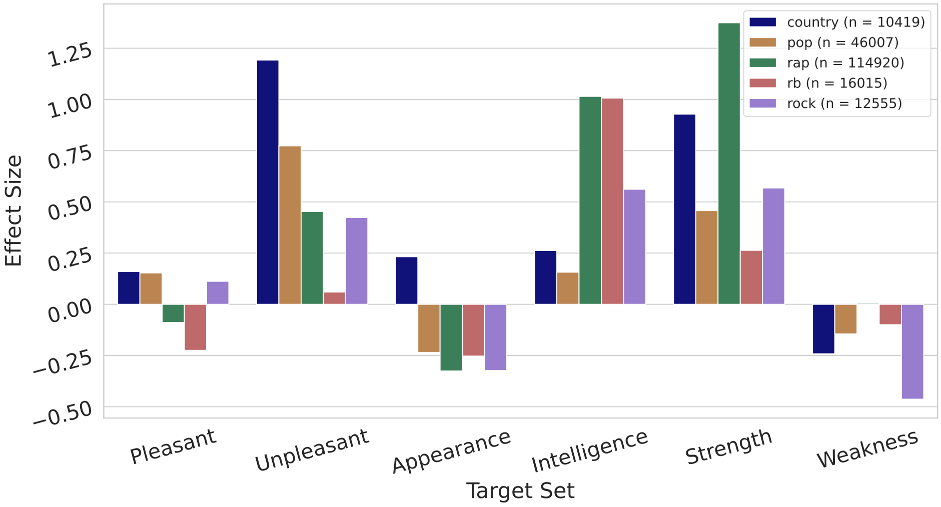


Figure 5: The SC-WEAT effect size of the target sets in each genre. A positive score indicates male bias, whereas a negative score indicates female bias, and n represents the number of word vectors for each genre.

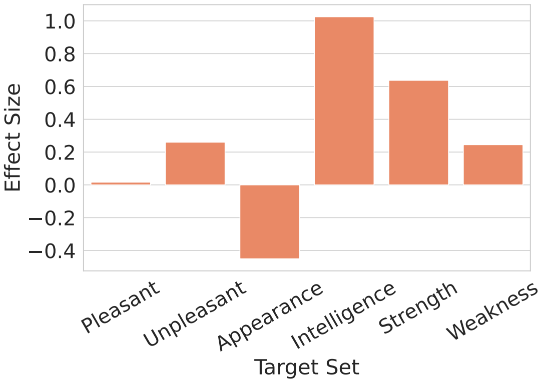


Figure 6: SC-WEAT score for the top topic: *"nigga_niggas_bitch"*. A positive score indicates male bias, whereas a negative score indicates female bias, and n is the number of word vectors.

all target sets, with Weakness words showing the strongest bias. These findings align with prior research by Rasmussen and Densley (2017), which observed that over half of the country songs analyzed reinforce stereotypical female gender roles and objectify women. This underscores the role of genre-specific contexts in shaping the gendered associations present in song lyrics.

Figure 8 reveals that Weakness words consistently exhibit a female bias across the three genres analyzed, aligning with our broader observation that women are more frequently associated with weakness. Notably, Intelligence words in country and R&B deviate from their average bias trends (see Figure 5), as these genres typically display a strong male bias overall yet show negligible scores for this specific genre.

The influence of genre-specific dynamics is further highlighted by the behaviour of Appearance words in Figure 8. While Appearance words display a male bias in R&B, they exhibit a female bias in pop, demonstrating how the same topic can exhibit divergent biases depending on the genre. These findings underscore the critical role of genre in shaping the gendered associations of recurring themes within song lyrics, emphasizing the need for a nuanced, genre-sensitive analysis to fully understand the interplay between thematic content and gender bias.

5 Conclusion

As a socio-cultural artefact, music offers insights into societal norms and biases, making it a valuable subject for computational analysis. This study leverages BERTopic, an advanced topic modeling technique, to identify thematic patterns and gender bias in song lyrics across five genres—country, pop, rap, R&B, and rock—over 70 years. Using SC-WEAT, we quantify gender bias within these themes and explore how biases vary across topics and genres. By addressing the intersection of music, culture, and societal norms, our findings reveal the gendered narratives embedded in song lyrics and their evolution over time.

We employ a stratified sampling strategy for BERTopic model training to ensure balanced genre representation. The most dominant topic, *"nigga_niggas_bitch"*, exhibits a high prevalence of misogynistic language and profanity, becoming particularly prominent in the 1990s despite the dataset spanning from the 1950s to the 2020s. In contrast, earlier dominant themes, such as *"tears_heart_wish"* and *"wonderful_sweeter_years_sweeter"*, primarily reflect

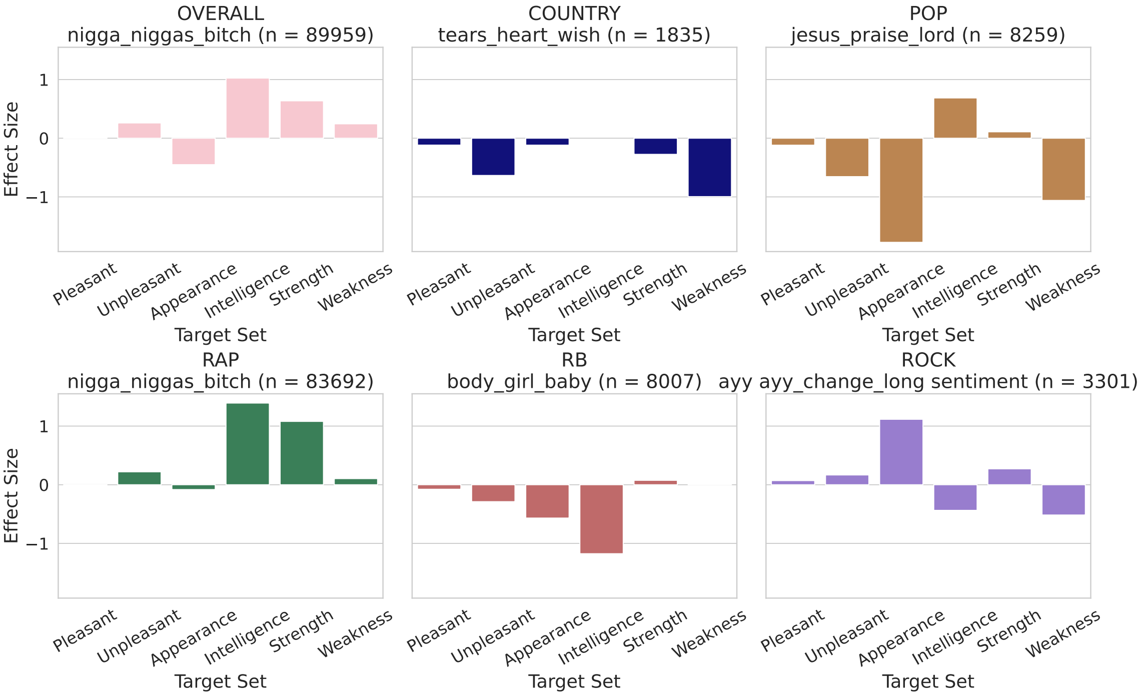


Figure 7: Comparison of SC-WEAT bias plots of the top topics in each genre. A positive score indicates male bias, whereas a negative score indicates female bias, and n is the number of word vectors.

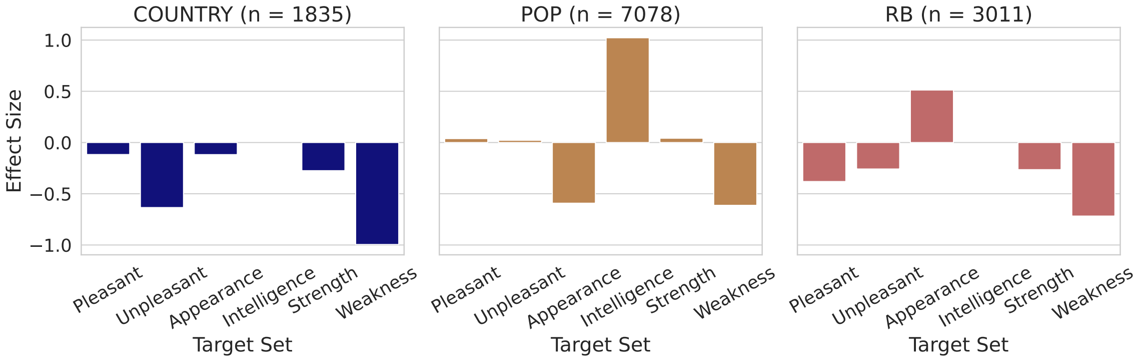


Figure 8: Comparison of SC-WEAT plots for *"tears_heart_wish"* in the country, pop, and R&B genres. A positive score indicates male bias, whereas a negative score indicates female bias, and n is the number of word vectors.

romantic and sentimental content. Over time, these themes shift toward heightened sexualization and explicit language, reflecting broader sociocultural and linguistic transformations in popular music, aligning with prior research on the increasing prevalence of sexualized and gendered language in song lyrics (Hall et al., 2011; Smiler et al., 2017b).

The SC-WEAT analysis further examines the trends of sexualization and profanity previously identified through topic modeling. The results reveal implicit gender bias in song lyrics, with Weakness and Appearance words showing a female bias, while Intelligence and Strength words exhibit a male bias. The female bias in Appearance words supports observations on the sexualization of women in music (Flynn et al., 2016; Hall et al., 2011; Rasmussen and Densley, 2017). The per-topic and per-genre analysis uncovers notable variations, with biases differing across themes and genres.

For instance, in the topic *"tears_heart_wish,"* bias scores vary across genres: country exhibits a female bias across all target sets, while Intelligence words in pop and Appearance words in R&B show a male bias. These results highlight the intersection of thematic content, genre, and gender bias, emphasizing the value of computational methods in analyzing sociocultural dynamics in song lyrics.

In conclusion, this study demonstrates the utility of integrating topic modeling with bias measurement techniques to analyze thematic structures in song lyrics and examine how these themes perpetuate implicit gender biases. By applying NLP methods to a significant sociocultural dataset, this work aligns with the growing demand in Digital Humanities and Social Sciences for tools that facilitate the analysis and interpretation of complex, non-standard textual data. Our approach highlights the potential of computational methods to address sociocultural questions, offering insights into how gender stereotypes are embedded in and perpetuated through lyrical content.

Limitations

Language Limitations: This study focuses exclusively on English-language songs, despite the multilingual content available on the Genius platform. Future research could expand to include songs in other languages, enhancing the scope and applicability of the findings.

Gender Classification: This analysis treats gender as binary, overlooking the spectrum of gender identities. Future research should explore the full spectrum of gender diversity in music for more inclusive insights.

BERTopic Modeling: A limitation of BERTopic, when applied to song lyrics analysis, is that it assigns a single topic per song, which does not account for songs that comprise different verses which may have different topics.

Race and Gender: In this paper, we look at the gender bias in lyrics independent of the race or gender of the artists, potentially neglecting their influence on the bias in the songs, especially in genres like rap. Future work could focus on integrating these aspects for a more detailed analysis of bias in music.

Addressing these limitations could significantly advance the field, offering an even more nuanced and comprehensive perspective on the intersection of music, culture, and societal norms.

References

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

Susan Alexander. 1999. The gender role paradox in youth culture: An analysis of women in music videos. *Michigan Sociological Review*, pages 46–64.

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

Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. 2018. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30.

Vincent P Bengio Y, Ducharme R. 2000. A neural probabilistic language model. In *NIPS*, volume 13, pages 932–938. MIT Press.

Lorenzo Betti, Carlo Abrate, and Andreas Kaltenbrunner. 2023. Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1):10.

Reihane Boghrati and Jonah Berger. 2022. Quantifying gender bias in consumer culture. *CoRR*, abs/2201.03173.

Reihane Boghrati and Jonah Berger. 2023. Quantifying cultural change: Gender bias in music. *Journal of Experimental Psychology: General*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Brook Bretthauer, Toni Schindler Zimmerman, and James H Banning. 2007. A feminist analysis of popular music: Power over, objectification of, and violence against women. *Journal of Feminist Family Therapy*, 18(4):29–51.

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

Kaytlin Chaloner and Alfredo Maldonado. 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.

Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 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.

Yihao Chen and Alexander Lerch. 2020. Melody-conditioned lyrics generation with seqgans. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 189–196. IEEE.

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

Maibam Debina Devi and Navanath Saharia. 2020. Exploiting topic modelling to classify sentiment from lyrics. In *Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, July 30-31, 2020, Proceedings, Part II 2*, pages 411–423. Springer.

Kevin Durrheim, Maria Schuld, Martin Mafunda, and Sindisiwe Mazibuko. 2023. Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, 62(1):617–629.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498.

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

Michael Fell, Elena Cabrio, Maroua Tikat, Franck Michel, Michel Buffa, and Fabien Gandon. 2023. The wasabi song corpus and knowledge graph for music lyrics analysis. *Language Resources and Evaluation*, 57(1):89–119.

Mark A Flynn, Clay M Craig, Christina N Anderson, and Kyle J Holody. 2016. Objectification in popular music lyrics: An examination of gender and genre differences. *Sex roles*, 75:164–176.

Cynthia M. Frisby and Elizabeth Behm-Morawitz. 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.

Lin Gan, Tao Yang, Yifan Huang, Boxiong Yang, Yami Yanwen Luo, Lui Wing Cheung Richard, and Dabo Guo. 2023. Experimental comparison of three topic modeling methods with lda, top2vec and bertopic. In *International Symposium on Artificial Intelligence and Robotics*, pages 376–391. Springer.

Klara Grönevik. 2013. The depiction of women in rap and pop lyrics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

P. Hall, Joshua West, and Shane Hill. 2011. Sexualization in lyrics of popular music from 1959 to 2009: Implications for sexuality educators. *Sexuality Culture*, 16.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Eirini Karamouzi, Maria Pontiki, and Yannis Krasonikolakis. 2024. Historical portrayal of greek tourism through topic modeling on international newspapers. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 121–132.

Kathrin Karsay, Jörg Matthes, Lisa Buchsteiner, and Veronika Grosser. 2019. Increasingly sexy? sexuality and sexual objectification in popular music videos, 1995–2016. *Psychology of popular media culture*, 8(4):346.

Florian Kleedorfer, Peter Knees, and Tim Pohle. 2008. Oh oh oh whoah! towards automatic topic detection in song lyrics. In *Ismir*, pages 287–292.

Louise Krasse. 2019. A corpus linguistic study of the female role in popular music lyrics.

Cyril Laurier, Jens Grivolla, and Perfecto Herrera. 2008. Multimodal music mood classification using audio and lyrics. In *2008 seventh international conference on machine learning and applications*, pages 688–693. IEEE.

Bingqing Liu, Kyrie Zhixuan Zhou, Danlei Zhu, and Jaihyun Park. 2023. Understanding gender stereotypes in video game character designs: A case study of honor of kings. 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 125–131, Tokyo, Japan. Association for Computational Linguistics.

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

Elissa Nakajima Wickham. 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, Tokyo, Japan. Association for Computational Linguistics.

Xuanlong Qin and Tony Tam. 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.

Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries.

Eric E Rasmussen and Rebecca L Densley. 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.

Andrew Smiler, Jennifer Shewmaker, and Brittany Hearon. 2017a. From "i want to hold your hand" to "promiscuous": Sexual stereotypes in popular music lyrics, 1960–2008. *Sexuality and Culture*, 21:1–23.

Andrew P Smiler, Jennifer W Shewmaker, and Brittany Hearon. 2017b. From “i want to hold your hand” to “promiscuous”: Sexual stereotypes in popular music lyrics, 1960–2008. *Sexuality & Culture*, 21(4):1083–1105.

Yi Yu, Abhishek Srivastava, and Simon Canales. 2021. Conditional lstm-gan for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1):1–20.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

A Appendix

A.1 Data Cleaning

We gather the song metadata from the WASABI corpus [4] and their respective lyrics information from the Genius Music Platform. Songs obtained from the Genius platform require preprocessing due to their unique format. Metadata associated with songs is typically enclosed within square brackets and embedded directly within the lyrical content. Additionally, the structure of the lyrics is generally preserved, resulting in entries that contain numerous newline characters. These characteristics may introduce challenges when parsing the data or preparing it for input into computational models, necessitating careful preprocessing to ensure consistency and usability. An example of the lyrics stored in the Genius dataset for “Love Story” by Taylor Swift:

> [Verse 1]
> We were both young when I first saw you
> I close my eyes and the flashback starts...
>
> [Pre-Chorus]
> That you were Romeo, you were throwing pebbles
> ...

A.2 Analysis of genre popularity across decades

Figure 9 presents a line chart illustrating the temporal evolution of genre popularity from the 1950s onward. In the early decades, country music demonstrates a higher relative prevalence compared to rap. However, a pronounced shift emerges in the 1990s, marked by a significant and rapid increase in the prominence of rap music.

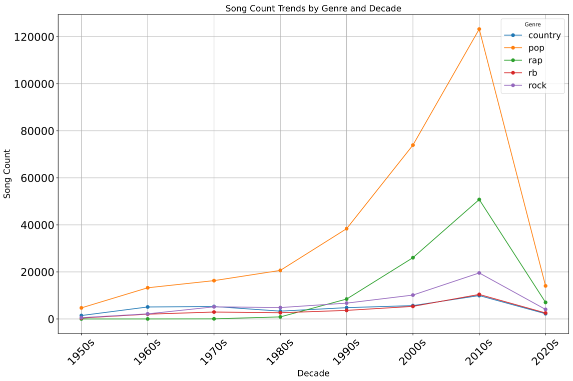


Figure 9: Genre trends over decades.

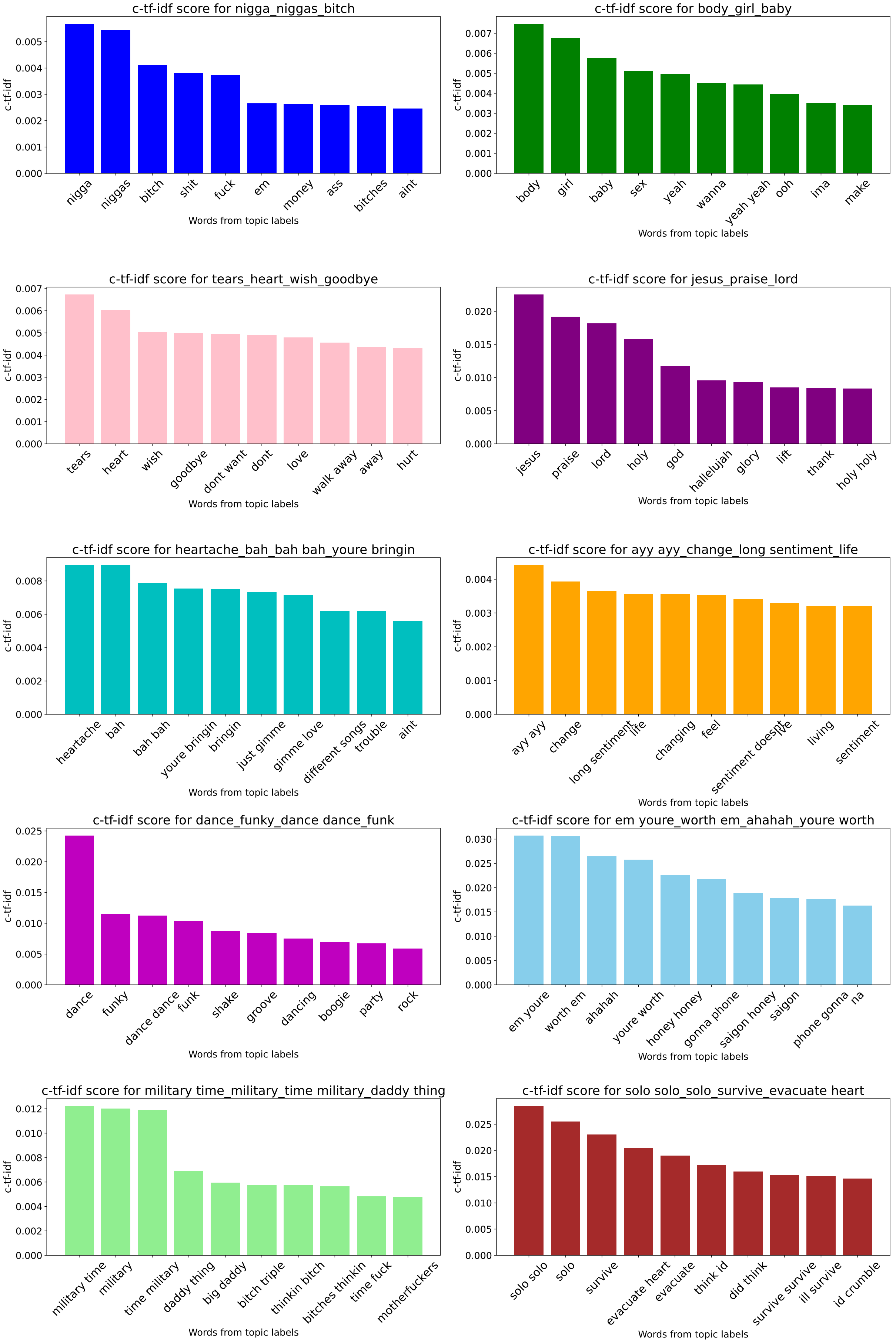
A.3 Initial BERTopic Model

The initial BERTopic model was trained on a randomly sampled subset of approximately 40,000 rows from the dataset. However, this approach resulted in an excessively high outlier rate, with over 50% of entries (approximately 27,000 rows) classified as outliers. This necessitated computationally intensive post-processing steps for outlier reduction, ultimately rendering the model suboptimal for integration into the final analytical pipeline. To address this limitation, we employed a stratified sampling strategy, selecting 107,000 rows balanced across musical genres for model training, followed by transformation on the entire dataset. This revised approach led to a substantial improvement in model stability and representational fidelity, reducing the proportion of outliers to just 1.5%. Consequently, this methodological refinement enhanced both the computational efficiency and the overall robustness of the topic modeling pipeline.

A.4 Topic Label Analysis Using c-TF-IDF score from Bertopic model

As shown in Figure 10, the topic labels are derived by selecting words with the highest c-TF-IDF scores, which are identified by the BERTopic model (Grootendorst, 2022). Unlike traditional TF-IDF, c-TF-IDF computes word importance at the cluster level rather than the document level (Ramos, 2003; Grootendorst, 2022). This method ensures that the most representative and distinguishing terms for each topic are highlighted, facilitating the interpretation of thematic structures within the dataset.

[4] https://github.com/micbuffa/WasabiDataset


c-tf-idf score for nigga_niggas_bitch

c-tf-idf score for body_girl_baby

c-tf-idf score for tears_heart_wish_goodbye

c-tf-idf score for jesus_praise_lord

c-tf-idf score for heartache_bah_bah bah_youre bringin

c-tf-idf score for ayy ayy_change_long sentiment_life

c-tf-idf score for dance_funky_dance dance_funk

c-tf-idf score for em youre_worth em_ahahah_youre worth

c-tf-idf score for military time_military_time military_daddy thing

c-tf-idf score for solo solo_solo_survive_evacuate heart

Figure 10: c-TF-IDF scores for words in the top 10 topics.

Target Set	Words
Pleasant	“friend”, “joy”, “wonderful”, “vacation”, “love”, “honest”, “honor”, “pleasure”, “loyal”, “family”, “peace”, “heaven”, “cheer”, “freedom”, “diploma”, “gentle”, “happy”, “paradise”, “diamond”, “laughter”, “sunrise”, “gift”, “health”, “rainbow”, “caress”, “lucky”, “miracle”
Unpleasant	“terrible”, “prison”, “divorce”, “war”, “poverty”, “sickness”, “abuse”, “tragedy”, “hatred”, “crash”, “accident”, “poison”, “nasty”, “awful”, “grief”, “disaster”, “stink”, “pollute”, “ugly”, “rotten”, “filth”, “failure”, “bomb”, “horrible”, “jail”, “kill”, “cancer”, “death”, “murder”, “evil”, “vomit”, “agony”, “assault”
Appearance words	“sensual”, “thin”, “handsome”, “feeble”, “bald”, “fashionable”, “slim”, “gorgeous”, “fat”, “plump”, “muscular”, “pretty”, “strong”, “weak”, “ugly”, “slender”, “homely”, “healthy”, “blushing”, “athletic”, “voluptuous”, “stout”, “beautiful”, “alluring”, “attractive”
Intelligence words	“intelligent”, “venerable”, “adaptable”, “reflective”, “thoughtful”, “resourceful”, “genius”, “logical”, “smart”, “astute”, “judicious”, “imaginative”, “intuitive”, “shrewd”, “ingenious”, “apt”, “precocious”, “inventive”, “analytical”, “inquiring”, “inquisitive”, “discerning”, “brilliant”, “clever”, “wise”
Strength words	“potent”, “bold”, “leader”, “strong”, “triumph”, “command”, “shout”, “winner”, “dominant”, “power”, “succeed”, “confident”, “dynamic”, “loud”, “assert”
Weakness words	“wispy”, “loser”, “failure”, “timid”, “lose”, “weak”, “weakness”, “shy”, “surrender”, “follow”, “fragile”, “withdraw”, “vulnerable”, “yield”, “afraid”

Table 3: List of target sets used for SC-WEAT analysis. These sets were chosen from the word sets curated by Betti et al. (2023), who compiled it from two different sources (Caliskan et al., 2017; Chaloner and Maldonado, 2019).

Attribute Set	Words
Female	“aunt”, “auntie”, “daughter”, “daughter-in-law”, “female”, “gal”, “girl”, “girlfriend”, “grandmother”, “grandmother-in-law”, “her”, “hers”, “lady”, “madam”, “mama”, “miss”, “mom”, “mother”, “niece”, “queen”, “she”, “sis”, “sister”, “wife”, “woman”
Male	“boy”, “boyfriend”, “brother”, “dad”, “father”, “father-in-law”, “grandfather”, “grandpa”, “guy”, “he”, “him”, “his”, “husband”, “king”, “male”, “man”, “nephew”, “papa”, “sir”, “son”, “son-in-law”, “uncle”

Table 4: List of attribute sets used for SC-WEAT analysis.