Exploring Online Music Discourse: A Gendered Lens On Social Media Music Artist Discussion

JoanMarie Leone
jleone5@binghamton.edu
Binghamton University
Vestal, New York, USA

Abstract

This study explores the perception and discussion of male and female music artists on Reddit and 4Chan, focusing on disparities in frequency of mentions and toxicity of posts. Using artist data sourced from two Kaggle datasets, I analyzed popular Western and K-pop artists across genres like pop and rap. Data was collected using a custom crawler, and toxicity scores were measured using the ModerateHatespeech API. Results show that female artists were more frequently discussed on both platforms, with this trend being particularly pronounced on 4Chan, where K-pop artists dominated the conversation. Additionally, toxicity scores revealed that female artists faced greater negativity compared to their male counterparts, with 4Chan exhibiting a broader range of toxicity levels than Reddit. These findings highlight the gendered dynamics of online discourse and the influence of platform-specific cultures on how artists are perceived and discussed.

1 Introduction

My project aims to explore how mainstream male vs female music artists and groups are perceived and discussed on the social media platforms Reddit and 4Chan. Women are often held to different standards than men in society, and this disparity extends to the music industry. Whether it's for their physical appearance, singing ability, song lyrics, or even personal life, it is said that female artists and groups face harsher criticism and judgment than their male counterparts. When it comes to song lyrics, singing about past relationships and sexual topics is criticized more when coming from women. Taylor Swift, one of the world's most popular female music artists, has been vocal about the public's sexist treatment towards her for years. Known for writing song lyrics about

past relationships, she talks about how male artists do the same thing all the time, but do not face the same type of judgment.

This paper focuses on popular music artists spanning multiple popular music genres. In particular, pop, rap, and K-pop (Korean pop) artists and groups make up the majority of the artists being collected. Artist names were taken from two online datasets from the online data science platform Kaggle. With the ongoing rise of K-pop in North America and across the world, it is important to include K-pop artists in this research [3].

To investigate these claims, I have collected data from Reddit's Music and popheads subreddits, as well as 4Chan's /mu/ board. I've set up a continuously running data crawler to begin collecting data from both platforms. I analyzed the frequency at which the different music artist were mentioned across Reddit and 4Chan, seeing that female artists tend to be more often discussed especially on 4Chan. I also used ModerateHatespeech to measure the toxicity of the posts regarding each of the artist's.

2 Background

Exploring the online discussion of male and female music artists is vital because it reflects broader societal attitudes and biases toward gender. Understanding these dynamics in social media discussions can provide valuable insights into how societal norms influence public perception and engagement with artists. Additionally, examining these patterns can shed light on how online platforms contribute to reinforcing or challenging traditional gender roles in the entertainment industry and beyond. To gain a broader and more inclusive collection of data, I have included K-pop artists into the data collection. K-pop's growing widespread influence, shown by its significant presence on platforms such as YouTube and Billboard [3], make it an important addition to studying the overall view of music artists on social media. K-pop artists, especially females, also have their own strict beauty standards that they try to adhere to [5] which could foster a lot of online discussion around their looks.

In my search, not much scientific research has been done regarding quantifying or analyzing the discussion of music artists on social media platforms. A previous study explores the preferences of promoters (those responsible for discovering new music talent and shaping artists' careers) and listeners, revealing that both groups exhibit a roughly equal

preference for male and female artists [4]. This could suggest that online social media discussions about artists should not show significant differences in the quantity of discussion based on gender. Another paper regarding commenting on online news sites discussing celebrities found that readers would sometimes aggressively comment on celebrities' weight and appearance even when the news articles had nothing to do with appearances [8]. This could suggest that the same type of behavior extends onto other social media platforms. There is limited research exploring the differences in online treatment of male and female artists. Given that platforms like Reddit, and especially 4Chan, are known for their open-ended and unfiltered nature, it is crucial to examine how users discuss female and male celebrities on these platforms.

3 Artist Dataset Creation

To begin creating the artists' datasets, I manually searched Kaggle, an online data science platform that offers various datasets, tools, and learning resources. Kaggle has a variety of datasets including those relating to music artists. My goal in the artists dataset is to encompass the most popular and relevant artists, including artists that are likely to be discussed on Reddit and 4Chan. In manually searching for datasets, the datasets needed to include the gender of the artists, since that is a main concern in my data collection.

3.1 popular_artists Dataset

The first online dataset I used in the creation of my own is "Spotify Artist Metadata Top 10k" which contained the top 10k artists from Spotify along with their sex and other details about the artist. The top 10k artists were obtained from Spotify and the metadata (including the artists set) was retrieved from the MusicBrainz database which stores information on music artists. This dataset had a column differentiating whether the artist was an individual person or a group. One shortcoming I noticed was that all groups, regardless of the gender of all the members, were automatically denoted 'mixed' under the gender column. To try and include at least some groups comprising members of all the same gender, I took the top 50 groups and manually assigned 25 of them male or female. Groups consisting of both men and women were excluded from the dataset. For my dataset, I took the top 100 male artists, the top 100 female artists, and the 25 groups. The artists in the dataset were ranked based on popularity, which is what I used to get the top artists in each category. One shortcoming of this dataset is that it was last updated four years ago which leaves a lot of new relevant artists out of the mix. To make up for this, I manually included artists nominated for a Grammy in 2025 who were not already on the list. This resulted in 18 new artists to the dataset (9 male and 9 female).

3.2 kpop_artists Dataset

In my search on Kaggle, I was able to find a Kpop dataset titled "Kpop Idol Instagram Followers". This dataset lists the 378 most followed K-pop idols along with information such as each artist's gender, music group (if applicable), stage name, and real name. The number of followers on Instagram, another social media platform, would be a good indicator of the most popular and talked about artists on these other social media platforms Reddit and 4Chan. From this dataset, I took the top 100 female and top 100 male artists based on the number of Instagram followers. The names of all the groups of the top artists were also taken into my dataset. The groups of all the top artist were also collected and assigned a gender based on their members gender. In Kpop, groups almost always consist of all the same gender so there was no conflict in assigning the groups their own gender.

3.3 Dataset Creation

Databases from Kaggle can be downloaded and accessed using the Kagglehub external library. The Python library pandas was then used to filter, sort, combine, and export the data to create my own datasets to use in my data collection. I took both Kaggle datasets and created two of my own datasets represented in a csv file. The first dataset titled popular_artists was created from the "Spotify Artists Metadata Top 10k" data, and the second dataset titled kpop_artists was created from the "Kpop Idol Instagram Followers" data.

Next, a function was created to parse a body of text for any occurrences of the artists' names. Individual artists are often referred to by just their first name, so for individual artists in the popular_artsits dataset, instances where only the first name of the artist is present are still detected. In the kpop_artists dataset, names were detected only if the full name was present. This approach was chosen for several reasons. First, most K-pop names are already a single word. Second, K-pop idols are typically referred to by their full stage names, as it is uncommon to use just part of their name. Unlike traditional American names with distinct first and last components, K-pop stage names often function as a cohesive identity, making partial detection less meaningful.

For Reddit posts, depending on whether the post was a comment or an original post, different fields were parsed for artist names. The two types of posts are differentiated from each other based on the 'kind' field or the beginning of the post id; Original posts have a post id beginning with 't3' while comments begin with 't1'. For original posts, the 'title' field of the JSON object, as well as the 'selftext' field (which represents the body of the post), were parsed. For comments, just the 'body' field was parsed. For all 4Chan posts, the 'com' field was checked for any artist names. This field represents the comment or body of a post. Posts were parsed for artists was they were being collected. For posts that were already collected before the implementation of the artist scanning, I

implemented a program to grab and update 200 entries from each database (one for the Reddit database and one for the 4Chan database) every two minutes. The process of updating these posts closely mirrors the approach used for incoming posts. Depending on the type of post, specific sections of the text data are extracted and scanned for any mentions of artists.

3.4 Dataset Problems

A couple of problems arose while identifying artists in posts. First, the first name detection of individual artists from the popular_artists dataset had a lot of individual artist stage names that did not follow the format of traditional first and last names. This would cause an overwhelming number of posts identifying an artist because it had the first part of an artist's name meanwhile the posts were not actually talking about the artist. One clear example was with the artist The Weekend. My artist scan function would identify any post with "The" in it as a post talking about the artist The Weekend. To fix this, individual artists that didn't work with first name detection had their type changed from 'person' to 'group'. The artist scan function only takes the first part of artists whose type was 'person', so now those artists labeled as 'group' will only be detected with their full stage name. Approximately 20 artists from the popular_artists dataset were changed to type 'group'. Further, during data analysis it was discovered that I missed the relabeling of certain artists which resulted about 10 artists being excluded from the study.

Another problem arose with artist stage names that only consisted of one or two letters. There were approximately 6 artist stage names from the kpop_artists dataset that were commonly falsely identified because their stage names were only one or two letters. Stage names like L, LE, and JR are commonly found in posts not related to the artist because of the nature of their names. Therefore, artists such as these were removed from the dataset. Two of the most popular Kpop artists' stage names' fell into this category: V and RM. To try to encompass these artists, their stage names were replaced with their first names in the search. While this would not fully encompass these artists to the same extent as their stage name, these artists are also identified by their first name and thus I would at least be able to include these artists in my dataset.

4 ModerateHatespeech

To measure the toxicity of the posts being collected, I used the ModerateHatespeech API. This API is designed to detect and classify harmful or abusive language, such as hate speech, by analyzing text and identifying potentially offensive or discriminatory content. The implementation of the ModerateHatespeech API was similar to the process used for scanning artist names. The same bodies of text that were

Figure 1. Combined top 20 artists based on frequency from Reddit and 4Chan
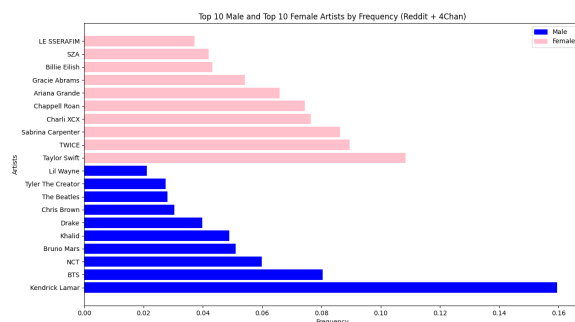
Figure 2. Combined top 10 male artists and top 10 female; artists based on frequency from Reddit and 4Chan

analyzed for artist mentions were also analyzed for toxicity using a hate speech detection function. This function used the ModerateHatespeech API to analyze the text and return a JSON object with the result. The API classifies the text as either 'normal,' indicating it is not hateful, or 'flag,' if the text is flagged as hateful. The JSON also contains a 'confidence' field, which gives a numerical value representing the API's confidence in its classification. This entire JSON object was stored in my database for later analysis.

5 Discussion

After continuously collecting posts for about a month, the number of posts mentioning each artist was analyzed and the frequency of the top artists' were calculated. You can see in Figure 1 that out of the top 20 artists, 13 of them are female. This is slightly more than the average which loosely supports the argument that women are discussed more frequently than men. This can be further seen in Figure 2 as the top 7-10th placed male artists all are discussed less frequently than the 10th placed female artist. While this is the case, you can also see that the top-discussed artist was Kendrick Lamar, a male artist. During the time of the data
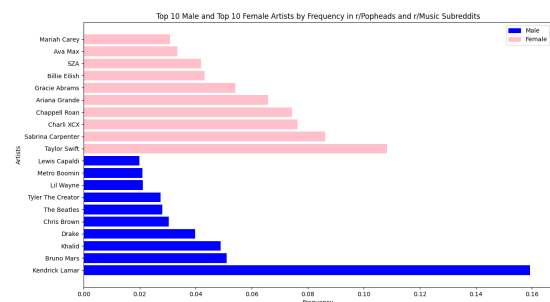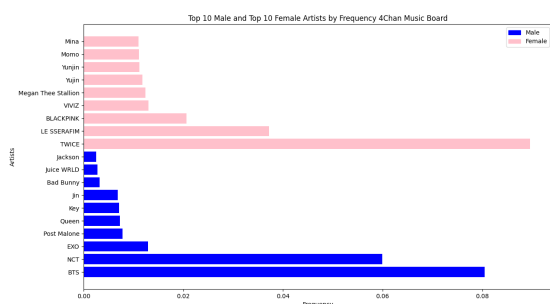
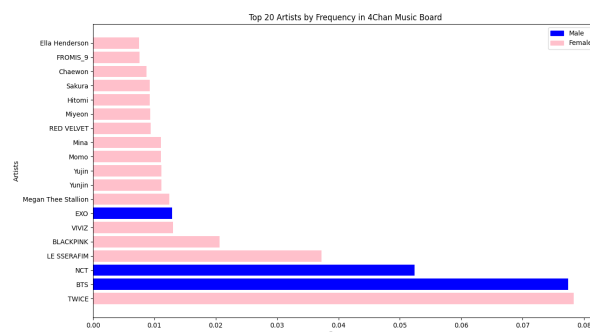**(a)** Top 10 male and female artists based on frequency from the Music and popheads subreddits
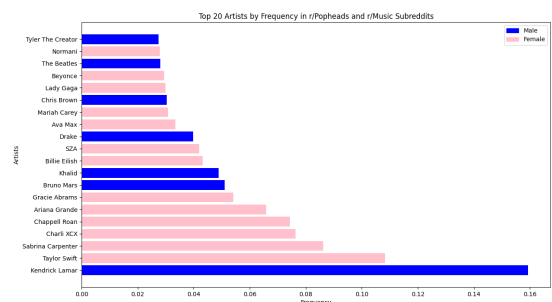
(b) Top 20 artists based on frequency from the Music and popheads subreddits

Figure 3. Reddit Top Artist Frequencies

(a) Top 10 male and female from 4Chan's /mu/ board

(b) Top 20 artists based on frequency from 4Chan's /mu/ board

Figure 4. 4Chan Top Artist Frequencies

collection, Kendrick Lamar, one of the most popular rappers, released a new album which is likely the reason for his high discussion rate. This could also show how the discussion of artists online correlates to real-world events and news surrounding these artists. It is also worth noting that 8 of the top artists are 2025 Grammy nominees which could further support this claim.

By separating the 4Chan and Reddit data, we can further analyze the different trends pertaining to each social media platform. Separating both you can still see a trend of the top female artists being mentioned a lot more frequently than their male counterparts.

5.1 Separating Reddit and 4Chan

Taking a deeper look into the Reddit data, we can see that from Figure 3b that just like in the overall evaluation 13 out of the 20 top-mentioned artists are female. All of the top artist also come from the popular_artists dataset which means that the top artists consist of popular western artists. The top-mentioned artist is Kendrick Lamar by a substantial margin. The distribution of the mentioned artists is skewed dramatically. Taking a look at the top 10 male and female

artists for Reddit, we can further see the discrepancy between Kendrick Lamar between other artists, especially his male counterparts. In figure 3a, we can see that the top 5 mentioned female artists are mentioned more than the all the top male mentioned artists excluding Kendrick Lamar.

Now looking at the 4Chan data, we can see a much bigger gap between the mentions of male and female artists. Figure 4b shows that out of the top 20 most mentioned artists, only 3 of them are male. This exemplifies the trend seen previously of female artists being discussed more than males. Another important feature to note is that all but 2 of the top 20 artists come from the kpop_artists dataset. This is almost the opposite of what was found in the Reddit data which had all of the top artists come from the popular_artists dataset. This could suggest that 4Chan users have different genre preferences when is comes to discussing artists. Also worth noting, Megan The Stallion the highest mentioned artist from the popular_artist dataset is currently collaborating with the top most-mentioned artsit TWICE (a Kpop girl group). This further supports the claim that 4Chan has a greater preference towards discussing Kpop especially female Kpop artists and groups. One possible explanation for this trend is that 4Chan
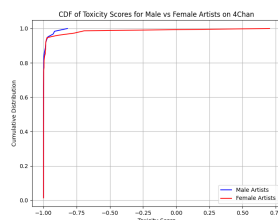
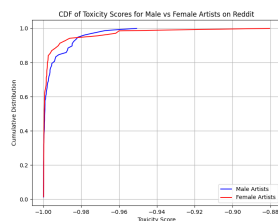**Figure 5.** 4Chan Male and Female CDF for Toxicity Scores

Figure 7. Number of comments per hour on r/politics from 11/01/24 to 11/14/24

Figure 6. RedditMale and Female CDF for Toxicity Scores

Figure 8. Number of submissions per day on r/politics from 11/01/24 to 11/14/24

users have an affinity to female Kpop idols and the beauty standards and cultural norms that these idols follow. Taking a look at the top 10 male and female artists for 4Chan we can see a much harsher cut off in frequency of artists mentions. The top 10 male artists decrease in frequency more dramatically than the female artists.

5.2 Toxicity Scores

To calculate the toxicity scores, I analyzed each post associated with an artist. If the post's class was labeled as "neutral," it was assigned a score of -1, and if it was labeled as "flag," it also received a score of -1. This score was then multiplied by the confidence level of the classification. Once the scores for all posts mentioning a specific artist were calculated, I averaged them to determine an artist-specific average toxicity score. Finally, I grouped all the scores for female artists and did the same for male artists, combining these values to create a CDF of toxicity scores based on gender. This was done separately for both the Reddit data and 4Chan data. The CDF's are represented in Figure 5 and Figure 6. Looking close at 4Chan's CDF we can see that the range of the toxicity scores goes from -1 to .75 meanwhile Reddit's only goes from -1 to -.88. This shows that the range of toxicity is a lot less in Reddit than 4Chan. We also see that in both Figures 5 and 6 the toxicity towards female artists was greater than that towards male artists.

6 Conclusion

Overall, analyzing the frequency and toxicity of discussions surrounding artists on Reddit and 4Chan, several notable trends emerged. Female artists were consistently discussed more frequently than male artists across both platforms, with
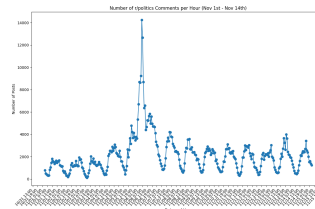
13 out of the top 20 mentioned artists being female. This pattern was even more pronounced on 4Chan, where only three male artists appeared in the top 20. This highlights a potential bias in the type of content users on these platforms engage with, reflecting differing preferences and possibly cultural factors unique to each community.

The stark contrast in the artists discussed on Reddit versus 4Chan further supports this observation. On Reddit, most of the top artists were from the popular_artists dataset, heavily favoring western artists. In contrast, 4Chan's discussions were dominated by female Kpop artists from the kpop_artists dataset, underscoring the platform's distinct preference for Kpop content. This suggests that platform-specific cultural dynamics and user interests play a significant role in shaping online discussions.

When examining toxicity, the results revealed that while Reddit had a narrower range of scores, both platforms exhibited higher levels of toxicity directed toward female artists compared to male artists. This trend aligns with broader concerns about how female figures in the public eye, including musicians, face disproportionate levels of scrutiny and negativity online. The significantly wider range of toxicity scores on 4Chan suggests a more volatile discourse environment compared to Reddit, which may reflect differences in platform culture and moderation practices.

These findings underscore how online platforms both shape and mirror cultural and social biases in the discussion of artists. Building on these insights, future research could explore the following questions: What underlying factors contribute to the disparity in toxicity levels directed toward male and female artists on Reddit and 4Chan? Why are female artists discussed more frequently than their male

counterparts? And what are the primary topics of conversation when social media users discuss male versus female artists?

[1] [2] [5] [7] [6] [8] [9]

References

[1] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–15.

[2] Jeremy Blackburn and Haewoon Kwak. 2014. STFU NOOB! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*. 877–888.

[3] Isabella Chen. 2023. Expansion of K-pop in the Global Market. (2023).

[4] Patricia L Donze. 2017. Gender and popular culture: a comparison of promoter and listener preferences for popular music artists. *Sociological perspectives* 60, 2 (2017), 338–354.

[5] Hiu Yan Kong. 2016. *The globalization of K-pop: The interplay of external and internal forces.* Ph. D. Dissertation. Hochschule Furtwangen.

[6] Haewoon Kwak and Jeremy Blackburn. 2015. Linguistic analysis of toxic behavior in an online video game. In *Social Informatics: SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers 6*. Springer, 209–217.

[7] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3739–3748.

[8] Gaëlle Ouvrein, Charlotte JS De Backer, and Heidi Vandebosch. 2018. Joining the clash or refusing to bash? Bystanders reactions to online celebrity bashing. *Cyberpsychology: journal of psychosocial research on cyberpspace.-Brno* 12, 4 (2018), UNSP–5.

[9] Jay Patel, Pujan Paudel, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2024. iDRAMA-Scored-2024: A Dataset of the Scored Social Media Platform from 2020 to 2023. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 2014–2024.