Sentiment Analysis By Gender on 2020's Top Charting Albums

Department of QTM, Emory University

QTM 340: Approaches to Data Science with Text

Niki Vasan and Rose Feng

Dr. Klein

December 9th, 2021

Introduction:

Music is everywhere. Literally. Streaming platforms like Spotify, Apple Music, and Youtube have made music from all over the world available at our fingertips at every time of day. Popular music is not just streamed by individual users; it is played in stadiums, shopping malls, grocery stores, TikTok audios, parties and more. The scope of cultural influence for popular artists has grown tremendously, as has the ability of their music to shape and reflect societal preferences on gender, race, language and more.

Given the sheer ubiquity of top-charting albums in the modern era, our motivation for conducting this project is to observe what music streaming behavior can tell us about societal biases and preferences, particularly in younger audiences. As with every other industry, the music industry has created and imposed its own gendered norms of what kinds of music are best performed by women and men. We want to quantitatively determine if there is a difference in the sentimental qualities of popular music authored by women as compared to men. Do people prefer listening to women whose lyrics tend to hold a more positive or negative sentiment? How would such a phenomenon relate to the new socio-cultural shift towards rejecting stereotypically feminine qualities of sweetness and docility?

We believe that this is an important topic to explore because as with many movements to empower minorities, oftentimes change can appear as a facade. In recent years, names such as Cardi B, Doja Cat and even Taylor Swift have received a lot of media attention for their desire to empower all forms of female expression and promiscuity through their music. However, we can see in even a quick glance of the Billboard charts that most of the top streaming albums are still created by male artists, some of whose music or musical genre could arguably run counter to

these narratives. Demystifying this phenomenon is a complex and evolving task, and we certainly don't claim to do so using this project. Instead, we view this project as a starting point to further understand, at a fundamental level, what kinds of musical content sell when sung by a woman versus a man.

Related Research:

Because we are interested in exploring the relationship between gender and music, we referred to three relevant studies that share similar themes and quantitative approaches with our project.

The first source we used is an academic paper (Anglada-Tort et al., 2019) written by Anglada-Tort, Manuel & Krause, Amanda & North, A., which analyzes popular music lyrics and musicians' gender over time using natural language processing and other machine learning techniques such as classification trees and linear mixed-effect models. These techniques can not only help researchers of this study identify the lyrical variables in the songs, but can also inform us of the methods they used to measure inequalities in gender distribution on top charting UK music from 1960 to 2015.

Although calculated popularity scores of the top female and male artists were very similar in the last four decades from late 1975 to 2015, female singers have been largely underrepresented compared to the men in the UK's music market. Learning about their results, we wanted to learn if there is a similar trend in women's music in the US market today. So, we chose to conduct our research within a more specific time period and in a different region, analyzing sentimental components of top songs in the US music market in 2020. Our project is closely

related to the research (Anglada-Tort et al., 2019) published by Anglada-Tort et al. because both studies explore the relationships amongst song lyrics, gender, and streaming behavior.

Another study that inspired us to focus on gender stereotypes in the music industry is a paper (Barman, 2019) written by Barman, M. P., Awekar, A., and Kothari, S. The central idea of their research is to better understand predefined social, gender, and career biases by computationally characterizing the lyrics styles of more than half a million songs over five decades. Based on the results, it is interesting to notice that even in the music market, men are more often associated with career-oriented goals, but women's names appear frequently in a family setting. This source helped us identify the second step of our analysis process, which is using sentiment analysis to determine the differing sentimental associations of men's music as compared to women's music. Although our project is similar to this research, we conducted our project in a different time span and scope since the popular albums in 2020 may reflect a more updated portrayal of streaming behaviors and trends in music in today's world.

To gain a deeper understanding of the methods we should use to analyze and compare the sentimental composition of song lyrics, we also referred to a data science project (Lytle, 2019) that targets general audiences who have a less technical background, authored by Brianna Lytle. Similar to the time span of this project, which studies Spotify's top musicians in 2019, we also focused our research on top artists in Billboard in recent years. What caught our eye was the quantitative approach that Lytle uses to analyze the similarities and differences amongst the lyrics in 2019's top albums. Similar to our method, Lytle's project also uses NLTK's Vader Sentiment Intensity Analyzer to compute the positive and negative sentiment scores and

visualizes the patterns by song. However, the thematic composition and the scope of her project are completely different from ours because Lytle solely compares the sentiments on an individual song level and focuses on studying the differences among four singers. In our project, however, we expanded the scope of Lytle's methodology to conduct a more in-depth sentiment analysis that measures the lyrical sentiment scores on both a song level and a verse level. We hoped that this would deliver more accurate and comprehensive insights on the differences between men's and women's lyrics. We will talk more about our analytical approach in the Process and Methods section.

Corpus Creation:

We created a dataset consisting of scraped albums from Genius, a website that provides lyrics for most released songs. Our albums were selected from Billboard's Top 200 of 2020 list. Because we were scraping our data by ourselves rather than using a preprocessed dataset, we wanted to ensure we were scoping our data generation process appropriately. We needed to achieve a balance between having sufficient data to run a successful sentiment analysis while also not taking up too much time or processing power. In the end, we decided to choose the top ten albums released by male artists and the top ten albums released by female artists. This resulted in 20 albums of 340 songs total.

The female artist albums we used were: *Folklore* by Taylor Swift, *When We All Fall Asleep, Where Do We Go?* by Billie Eilish, *Lover* by Taylor Swift, *Over it* by Summer Walker, *Manic* by Halsey, *Chilombo* by Jhene Aiko, *Hot Pink* by Doja Cat, *Cuz I Love You* by Lizzo, *Dont Smile At Me* by Billie Eilish, *Chromatica* by Lady Gaga. The male album artists we used were: *Hollywood's*

bleeding by Post Malone, *My Turn* by Lil Baby, *Please Excuse Me For Being Antisocial* by Roddy Ricch, *Fine Line* by Harry Styles, *Eternal Atake* by Lil Uzi Vert, *Shoot For The Stars Aim For The Moon* by Pop Smoke, *After Hours* by The Weekend, *Legends Never Die* by Juice WRLD, *What You See Is What You Get* by Luke Combs, *YHLQMDLG* by Bad Bunny.

So what does our data look like? We chose not to keep our data in a traditional structured csv format as we were anticipating needing to pass in documents (text files) to our sentiment analyzer. Instead, we have two folders, one for female artists and one for male artists. Within each folder, there are ten nested subfolders that represent each of the chosen albums for that gender. In each album folder, the lyrics for each song are saved as a text file whose title includes the artist name and song title for easy retrieval (e.g. Juice-wrld-wishing-well.txt).

Now, you may be wondering how we got our data in this format in the first place. Genius isn't a music streaming platform that necessarily stores songs in album format, and people typically use the website just to find lyrics of individual songs. This made our task a bit more difficult. We ended up using a compilation of functions from our in-class exercises with the API, aided by the *lyricsgenius* python library that helped us query the API by album. The result was a lengthy function that got us the data in five steps:

1. Use the *lyricsgenius* library to get json file of each album's track information
2. Create a dictionary of song titles and album artists by iterating through the json file of track information
3. Iterate through the data in Step 2 to create a dictionary of songs and their respective lyrics urls

4. Query the contents of the lyrics url for each song and clean the lyrics of any unnecessary new lines or characters
5. Write the lyrics to a ".txt" file with the appropriate title and save them to the album directory

Our last step was to call this function (titled *getalbumlyrics*) for each album-artist combination. Once we did this, our corpus was complete.

Process and Methods:

Now that we have gotten our corpus, which consists of all the song lyrics of the top 10 albums created by male artists and top 10 albums created by female artists in 2020, we can start conducting our sentiment analysis! To ensure the accuracy and viability of our sentiment analysis results, we first tried to determine the appropriate unit of analysis for our function input, since sentiment scores can be vastly different when they are computed by verse or by song. We ultimately decided on two approaches:

(1) Compute verse-level compound sentiment scores of each song and aggregate them using the means of verse-level scores to create one compound score for the song. Aggregate these averaged values of each song to create the overall album scores and compare them in the next steps.
(2) Calculate the compound sentiment score on each song and use the number as a basis for creating album-level aggregations, which will then be used for further analysis during the comparison process.

To test out which option works better, we started by using Talor Swift's song *August* in the *Folklore* album. However, even after we removed the outliers, the average of verse-level sentiment scores of *August* is: 0.49, which is drastically different from the song-level sentiment score 0.99. Because the results were not very telling, we decided to do the same thing for all the songs in the *Folklore* album.

	Verse-Level	Song-Level	Absolute Difference
peace	-0.590100	-0.5022	0.087900
mad-woman	-0.341029	-0.9862	0.645171
mirrorball	0.415800	0.8566	0.440800
the-last-great-american-dynasty	0.464533	0.9825	0.517967
hoax	-0.515240	-0.9825	0.467260
betty	0.183978	0.9572	0.773222
the-1	0.480243	0.9808	0.500557
this-is-me-trying	-0.102214	-0.9012	0.798986
invisible-string	0.326950	0.9470	0.620050
my-tears-ricochet	-0.329760	-0.9813	0.651540
cardigan	0.537412	0.9867	0.449288
august	0.328967	0.9905	0.661533
epiphany	0.348740	0.9081	0.559360
seven	0.582725	0.9908	0.408075
illicit-affairs	-0.235620	-0.8987	0.663080
exile	0.027250	0.2981	0.270850

Figure 1. Comparing verse-level and song-level sentiments in Taylor Swift's *Folklore*

As shown in Figure 1, the absolute differences between verse-level and song-level sentiment score are very high, meaning there is a significant discrepancy between the two approaches. Since we couldn't decide on which approach is more reliable at this point, we conducted

another calibration test using a human labeling approach to test out whether the song-level or verse-level sentiment scores are closer to our human labeled sentiments of song lyrics.

Human labeling is a method in which we label the sentiments of songs we are familiar with on a 5 point categorical scale, which is associated with a continuous interval ranging from -1.0 to 1.0.

Specifically, the categorical labels we used are:

Strongly negative -> Moderately negative -> **Neutral** -> Moderately positive -> Strongly Positive

And the quantitative scale of each of the categorical labels is:

[1,−0.6)−>[−0.6,−0.2)−>**[−0.2,0.2)**−>[0.2,0.6)−>[0.6,1.0]

By computing the absolute differences between our human labels and computed sentiment scores on a verse level and on a song level, we can decide on the appropriate unit that yields a more accurate and valid sentiment score. We used five songs from the male albums and five songs from the female albums for this calibration test.

	Song-Names	Song-Artists	HL Class	Song-Level-Scores	Verse-Level-Scores	Absolute-Difference	Accurate-Y-N
0	Hollywood's Bleeding	Post Malone	Strongly Negative	0.7503	-0.244050	0.506250	N
1	Can't Die	Juice WRLD	Moderately Negative	0.9898	0.417250	0.572550	N
2	For the Night	Pop Smoke	Neutral	-0.9979	-0.914933	0.082967	N
3	High Fashion	Roddy Ricch	Moderately Positive	0.9482	0.109150	0.839050	Y
4	Watermelon Sugar	Harry Styles	Strongly Positive	0.9502	0.203763	0.746438	Y
5	When the Party's Over	Billie Eilish	Strongly Negative	0.9964	0.452560	0.543840	N
6	Playing Games	Summer Walker	Moderately Negative	0.9180	0.436200	0.481800	N
7	Juicy	Doja Cat	Neutral	0.9977	0.887420	0.110280	N
8	Happiness Over Everything (H.O.E)	Jhene Aiko	Moderately Positive	-0.9879	-0.801300	0.186600	N
9	Soulmate	Lizzo	Extremely Positive	0.9999	0.793844	0.206056	Y

Figure 2. Comparing verse-level and song-level sentiments with human labels

As shown in Figure 2, we compared song-level and verse-level results with our human labels. We labeled the accuracy as Y if at least one of the song-level and verse-level scores match the human label. However, only 30% of the sentiment scores fall into our labeled category..

The differences we observed between the computed results and human-labeled results are so big that we failed to determine the appropriate unit for our sentiment analysis using this approach as well.

Given that we failed to determine the unit of analysis for our sentiment analyzer, we understood that our analytical approaches likely left a lot of room for error. As a resolution, we decided to continue with using both the song-level and verse-level sentiment model on all of our albums to analyze gender based disparities, attempting to avoid further inaccuracies in the final results. We will discuss the potential errors and ways to improve our sentiment model in the Next Steps section.

Results

As mentioned before, since both the absolute comparison and the human labeling approach of determining the unit of analysis yielded little results, we decided to incorporate both verse-level and song-level sentiment models into our final analysis. We did this with the knowledge that the verse-level score is likely to be more accurate since the scores tend to be less skewed towards a given extreme which is likely due to the multiple averages used to calculate the result. Figure 3 below is the data table of results for the female albums.

	Album	Verse-Level-Average	Song-Level-Average
0	folklore	0.154577	0.224569
0	WHENWEALLFALLASLEEPWHEREDOWEGO	0.172262	0.215015
0	Lover	0.126749	0.201129
0	OverIt	0.353774	0.341845
0	Manic	0.169664	0.099992
0	Chilombo	0.236165	0.187667
0	HotPink	0.349611	0.620627
0	CuzILoveYou	0.326391	0.367350
0	dontsmileatme	-0.141817	-0.435900
0	Chromatica	0.089364	0.167908

Figure 3. Sentiment scores of women's albums

The score we are aggregating here is the compound score produced by the VADER sentiment analyzer from the NLTK library. This is a score that ranges from -1 to 1, with a score of -1 indicating an extremely negative sentiment and a score of 1 indicating an extremely positive sentiment. In the case of the verse-level-average column, each verse in a song is given a score; that score is then averaged over the song, then over the album, then over all of the albums in a given gender (we assume the gender is binary in this case). In the case of the

song-level-average, the entire song is passed in as an input, eliminating the first step in the process outlined above.

If we take a look at the verse-level average scores, we can see that most of the scores lie near 0, which indicates neutrality. The overall average of the verse-level scores for female albums was 0.15, which is just above neutral, with a slight positive tilt. The average song-level sentiment score of the female albums was 0.32, which is not much higher.

Now, we can take a look at the results for the male albums.

	Album	Verse-Level-Average	Song-Level-Average
0	HollywoodsBleeding	-0.123975	-0.089721
0	MyTurn	-0.361158	-0.499138
0	PleaseExcuseMeforBeingAntisocial	-0.169493	-0.177093
0	FineLine	0.134158	0.328936
0	EternalAtake	0.036813	-0.391412
0	ShootfortheStarsAimfortheMoon	0.146611	0.237026
0	AfterHours	-0.084088	-0.218371
0	LegendsNeverDie	0.030895	-0.123156
0	WhatYouSeeIsWhatYouGet	0.345420	0.596062
0	YHLQMDLG	-0.443686	-0.946820

Figure 4. Sentiment scores of men's albums

Based on Figure 4, we can see right off the bat that the men's sentiment scores on both a verse level and a song level tend to be more negative than the women's scores. However, they also tend to be concentrated close to 0. The average verse-level sentiment score of the male albums is -0.05, and the average song-level sentiment score -0.13, which again, is not significantly more negative on a scale from -1 to 1.

The differences by gender can be better visualized in bar charts, as shown in Figure 5 and Figure 6.

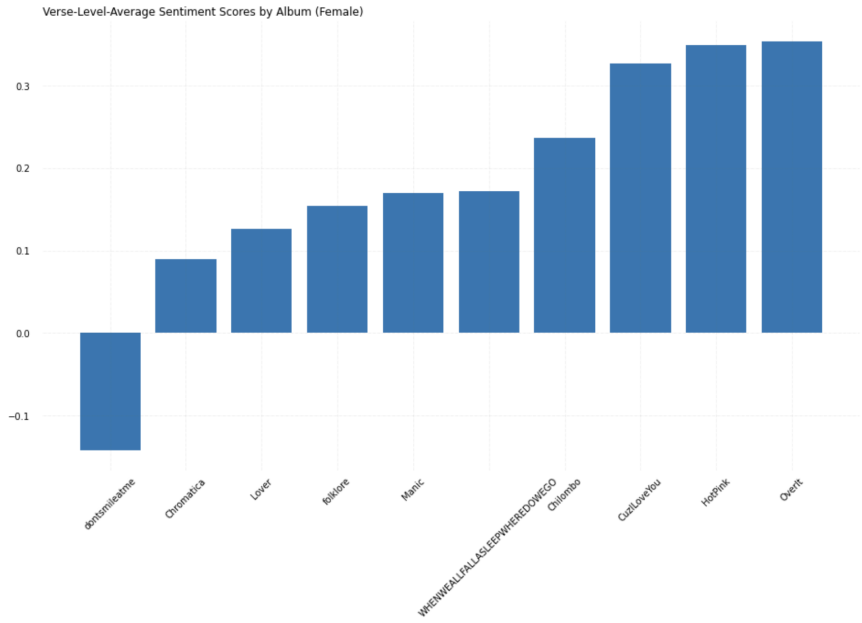

Figure 5. Average verse-level sentiment scores by women's albums

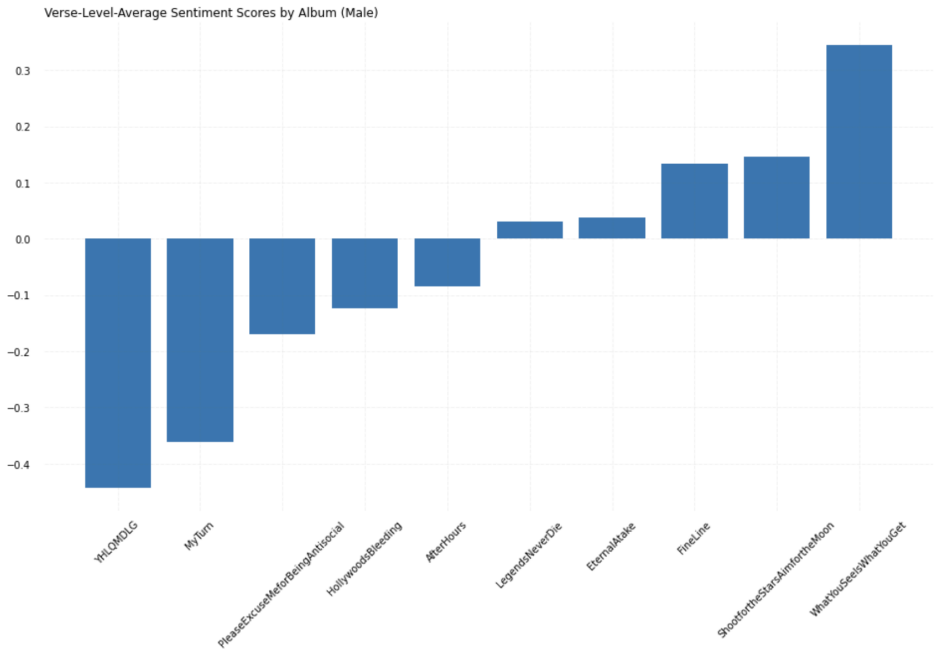

Figure 6. Average verse-level sentiment scores by men's albums

Here we can see the differences in the verse-level-average distribution between male and female albums. The chart on the left represents female albums, while the chart on the right represents male albums. As we can see above, the only female album with a slightly negative sentiment score is *Don't smile at me* by Billie Eilish, which is an explicitly sad, somewhat haunting album that is characteristic of the artist's persona and musical style. However, the results of the albums, including Billie Eilish's other album *WHEN WE ALL FALL ASLEEP, WHERE DO WE GO?*, all had positive sentiment scores.

The male albums, on the other hand, have a very different distribution. Half of the albums have positive sentiment scores, while the other half have negative sentiment scores. The negative scores are also more extreme than the positive ones. We do have to note, however, that *YHLQMDLG* by Bad Bunny (the leftmost bar) is an album that is entirely in Spanish. We did not realize this prior to completing the project, but the VADER sentiment analyzer only provides accurate sentiment analysis on English texts, meaning this score should be discounted.

Discussion

From these preliminary results, we can clearly see that the sentiment analysis did not yield particularly telling results. The verse-level and song-level average sentiment scores for both genders are close to 0, meaning they yielded mostly neutral results. This could either be due to the type of sentiment analyzer we used (i.e. we chose to use NLTK's VADER over BERT's HuggingFace Sentiment Analyzer, for example) or due to errors in our methodology.

Unfortunately, we are a bit too far in to redo our entire analysis with a different model. What we can do, however, is dive a little deeper into our methodology to see how we can improve our approach.

Given our baseline knowledge of each artist's genre and general musical style, we know that many of these results are headed in the right direction. For example, Lizzo's *Cuz I Love You* is an album that gained popularity due to its upbeat and positive message of self-love and female empowerment. Its sentiment score was positive (0.32), but not as positive as you might expect. Here is a closer look at the song-level scores of the album.

	Verse-Level	Song-Level
likeagirl	0.363786	0.9718
soulmate	0.793844	0.9999
cuziloveyou	0.427233	0.9676
heavenhelpme	0.490671	0.9994
betterincolor	0.717025	0.9990
tempo	-0.481043	-0.9977
lingerie	0.401250	0.9696
exactlyhowifeel	0.095617	0.9889

Figure 7. Sentiment scores of Lizzo's *Cuz I Love You*

We can see that most of these songs have actually very high verse-level sentiment scores, but the aggregate sentiment score of the entire album was dampened by the presence of one or two songs (*tempo* and *exactly how I feel*). The median score of this album is 0.41, which is about 0.1 units higher than the average and arguably a bit more reflective of the true album

sentiment. Clearly, this is just one example, but it begs the question of whether the mean was the appropriate statistic to use in this case when aggregating the sentiment scores.

Another broader question we asked ourselves after completing this project is whether using a sentiment analyzer on song lyrics will ever yield results that coincide with the listener's experience of hearing the album. Especially in light of increasingly popular genres like rap that have a fast lyrical pace or the increased exposure to music of different languages and cultures, many people's sentimental experience with music is more so attached to the musical qualities of the song rather than a thorough understanding of the lyrics. This could explain why our human labeling classification attempt was not successful; we labeled songs on a spectrum according to our experience *listening* to them, while the sentiment analyzer was solely looking at the textual composition of the song.

This becomes relevant as we make our way back to our original purpose for conducting this analysis in the first place - to determine what kinds of musical content sells when performed by a woman as compared to man. To answer this question holistically, we might want to find some way to account for the musical composition of songs (maybe partitioning by genre, for example).

For right now though, we can only conclude that songs in the most streamed albums from female artists are slightly more positive in nature, while the lyrics in men's albums are slightly more negative. Although the differences are relatively subtle, they are still telling about the relationship between streaming behavior and gender biases. On average, audiences prefer to listen to songs with a more positive sentiment from female artists than they do from male

artists, which tends to reinforce gendered expectations of how women should present themselves as compared to men.

Conclusion and Next Steps:

The results of our project are probably not the most revolutionary in the field of text analytics, but they do provide preliminary findings on the sentimental disparities in women's and men's lyrics. On average, people tend to prefer listening to music from women that has a more positive sentiment, but this standard does not hold for men. The fluidity in men's sentiment scores is congruent with their overrepresentation in the Billboard Top 200, meaning people may prefer listening to music with a sentimental variety. So the top 10 men's albums, which have more varied sentiment scores, significantly attracted more listeners, dominating Billboard's top album chart in 2020. This explanation could also potentially account for a similar phenomena in the U.K., as noted by Anglada-Tort et al. in their paper (Anglada-Tort et al., 2019), where male artists have dominated the top song charts from 1960 to today.

Or, people might just prefer listening to men's music in general. It tends to be more socially acceptable for women to listen to stereotypically masculine music than for men to listen to stereotypically feminine music, so further work should be done to dig deeper into this data.

As some next steps, we would like to refine our sentiment analysis methodology to better account for language differences and sentimental extremity in songs (so the scores are less neutral). We also think it would be interesting to conduct a regression analysis on the correlation between sentiment scores of lyrics and song popularity to study the factors that contribute to the gender inequality in popular music. Lastly, we hope to implement an in-depth

thematic analysis and a genre-based analysis to identify the topical compositions of song lyrics and compare them based on different genres, such as hip pop, folk, and pop music. By doing this, we hope to add nuance to our results and determine more factors that may influence the music streaming behaviors of audiences.

References:

1. Anglada-Tort, Manuel & Krause, Amanda & North, A.(2019). *Popular music lyrics and musicians' gender over time: A computational approach. Psychology of Music*. 49. 426-444. 10.1177/0305735619871602.
2. Barman, M. P., Awekar, A., & Kothari, S. (2019, July). Decoding the style and bias of song lyrics. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1165-1168).
3. Lytle, B. (2019, December 31). *Using natural language processing to analyze Spotify 2019 top global artists*. Medium. Retrieved November 11, 2021, from https://briannalytle7.medium.com/using-natural-language-processing-to-analyze-spotify-2019-top-global-artists-e5449d8b5133.