Were you really born in the wrong generation?

Emre Kurt

May 13, 2022

Abstract:

Conversations about music with people often include a mention of dislike of modern music, citing the music of their generation by a number of metrics. For this project, the focus is on the lyrical content of the songs. This project attempts to compare and contrast the lexical composure of each generation's most popular artists and their songs. Such comparisons include lexical diversity, sentiment, similarity, and word choice. Popularity of an artist is determined from the list of top selling musicians per decade. A song was chosen for each artist spanning their entire career, and merged with the songs of artists within the same time period. For example, The Beatles and Bob Dylan were placed in the same dataset as a representation of Baby Boomer songs. From there, each generation's songs had their R values (total unique words over total words) calculated, their most commonly used words displayed, their lyrical complexity compared, their emotion and sentiment analyzed, and similarities found. As a result, no significant differences between each generation's lyrical diversity was found, save for the Millennials who displayed the highest R values. There was also a noticeable emotional sentiment difference, where Millennials and Gen X showed to have a higher ranking in displays of negative sentiment than the other generations. This could partly be attributed to the popularization of alternative rock and hip-hop, whose lyrics are often about violence and negative topics. The Millennial generation shared the most similarity with Generation Z, which could be attributed to the rise of the Internet making cross-generational music sharing easier.

Introduction:

Imagine a hypothetical situation where you are talking about music with someone, and they mention that they do not enjoy modern music. Chances are that they will utter this exact sentence:

"I was born in the wrong generation." [1]

The sentiment behind this sentence is a longing for being able to experience the music of yesteryear, where in their opinion, the quality of production and lyricism far surpasses that of today's music. This negative sentiment is seen when comparing, say, Eminem versus Playboi Carti, the latter being criticized for his use of mumble rap [2]. Upon thinking of this conversational phenomenon, this question arises: is the music of the past actually better than that of the present? That is, are the words they used more intellectually provocative or creative?

Conversations surrounding art are a difficult path to tread. The enjoyment of art is known to be a subjective experience, where some find metal grating and abrasive while others find pop to be the same way. The means of measuring artistic skill have been deliberated through centuries of curation and discussion, leading to previously unknown artists being shown the recognition they deserve. Unfortunately, there can also be artists that are unfairly treated with the same eye of criticism. To combat this, the age old euphemism works well: "beauty is in the eye of the beholder".

I feel it appropriate, as a student who has taken a course in natural language processing, to add to the euphemism. "Beauty is in the eye of the beholder, but lexical diversity is in the hands of the data scientist."

That brings us to the hypothesis that this paper attempts to disprove. By taking each generation's most popular artists and the lyrics they use, can we find a difference in lexical diversity and sentiment?

Methods and results:

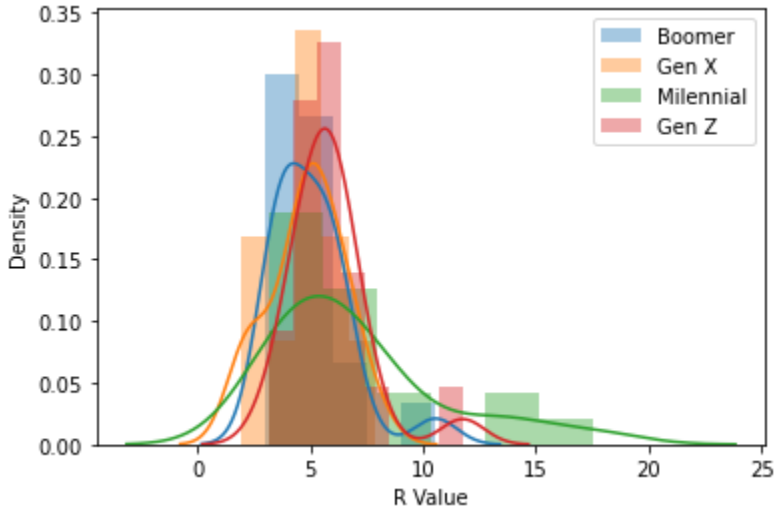
A popular artist is synonymous with artists who sell the most records. The rankings of each artist by sales per decade was found online [3]. A total of five artists per decade was chosen for each generation. The artists were not chosen by exact ordered ranking, but by subjective selection based on the list. For the Baby Boomers (born 1946 to 1964), The Beatles, Bob Dylan, Pink Floyd, Elvis Presley, and Led Zeppelin were chosen. For Generation X (born 1965 to 1980), Michael Jackson, Madonna, Queen, Nirvana, and Metallica were chosen. For Millennials (born 1981 to 1996), Eminem, Linkin Park, U2, Britney Spears, and Beyonce were chosen. For Generation Z (born 1997 to 2012): Drake, Taylor Swift, The Weeknd, Bruno Mars, Adele were chosen. Artists were not chosen based on their respective genres, because their popularity alone gives us a large sample space of people that listen and know about them.

Lyrics were scraped from AZLyrics.com [4], a popular website that contains lyrics for artists' songs. Songs were chosen on a stepped iteration (i.e. every nth based on the catalog length). This is to ensure an accurate representation of an artist's lyrical ability and to prevent any outlier albums. The reason all of each artists' songs were not chosen was because of computational and time limitations, as pulling every song's lyrics would take days. The functions (mainly using Beautiful Soup [5]) to pull the lyrics were largely my own, with portions taken from Stack Overflow forums and Medium articles.

Four songs from each artist were scraped, tokenized, had their stop words and punctuation removed (based on NLTK's parameters [6]), and placed in DataFrames respective to their artist, and then their generation.

The first metric measured was the R value, which is the number of unique tokens (words) divided by the square root of the total number of words in each song. This metric measures lexical diversity. The square root is to account for the length of the songs, which might be an issue with rap music. This value was calculated for each song in their respective generational DataFrame.

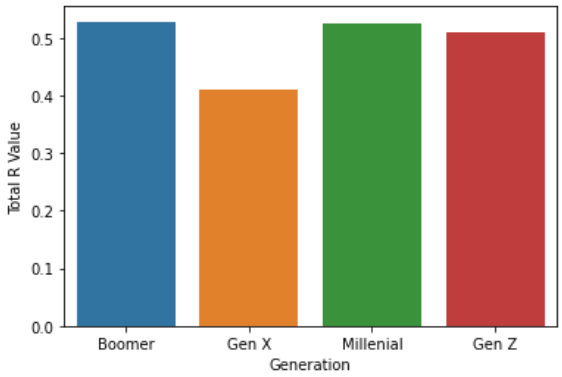
Distribution of R Values by generation



The distribution of R values for each generation shows no significant difference, except for the Millennial generation, which follows a slight right-skewed distribution. This means songs belonging to the Millennial generation often show more lexical diversity than the others'.

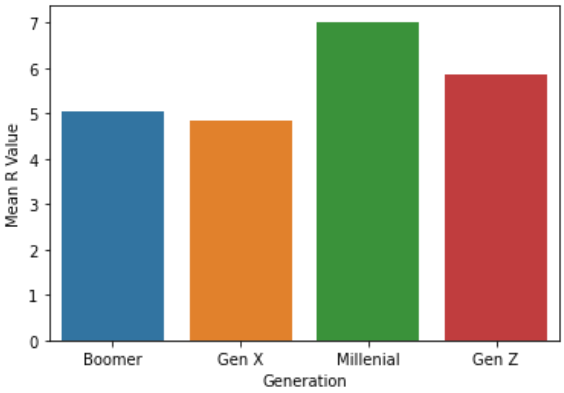
Then, the generation-wide R value was calculated. This merged all songs in each generation and calculated the R value based on the large concatenated string.

Generation-wide R Values

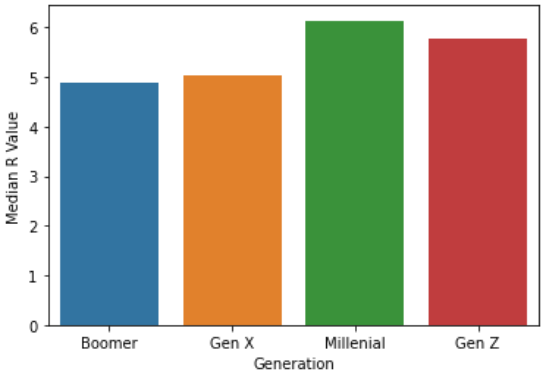


The R value was looked at twice more:

Mean R Value



Median R Value

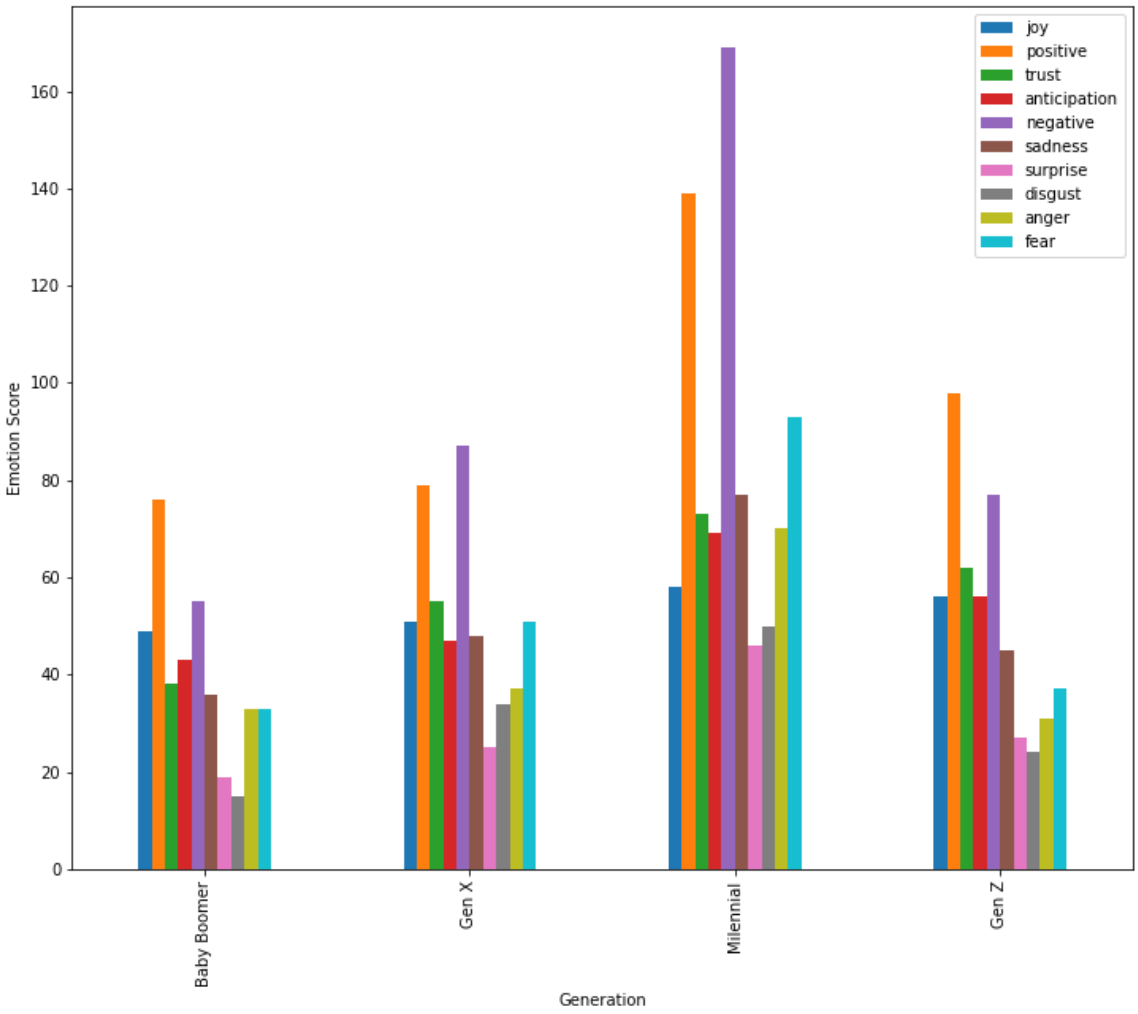


Out of these metrics, the Millennial generation's lyrics seem to have the highest lexical diversity, followed by Gen Z, then the Baby Boomers. This could partly be explained by the introduction of hip hop as a popular genre of music in the 90s, which categorically has more unique words in its lyrics.

The next metric calculated was the Flesch Reading Ease Score [7], where a lower score accounts to less readability. The formula uses total words, sentences, and syllables. The least readable generation was the Millennials, followed by Gen Z, then Gen X, then with Baby Boomers being the most readable.

Next, I used the NRC Word-Emotion Association Lexicon [8] to get a basic idea of emotional sentiment for each generation. The emotions accounted for are 'joy', 'trust', 'anticipation', 'sadness', 'surprise', 'disgust', 'anger', and 'fear'. Positive and negative sentiments can also be found using this package. Each generation was assigned a value for each emotion above based on their words and the emotional mapping in NRC's dictionary.

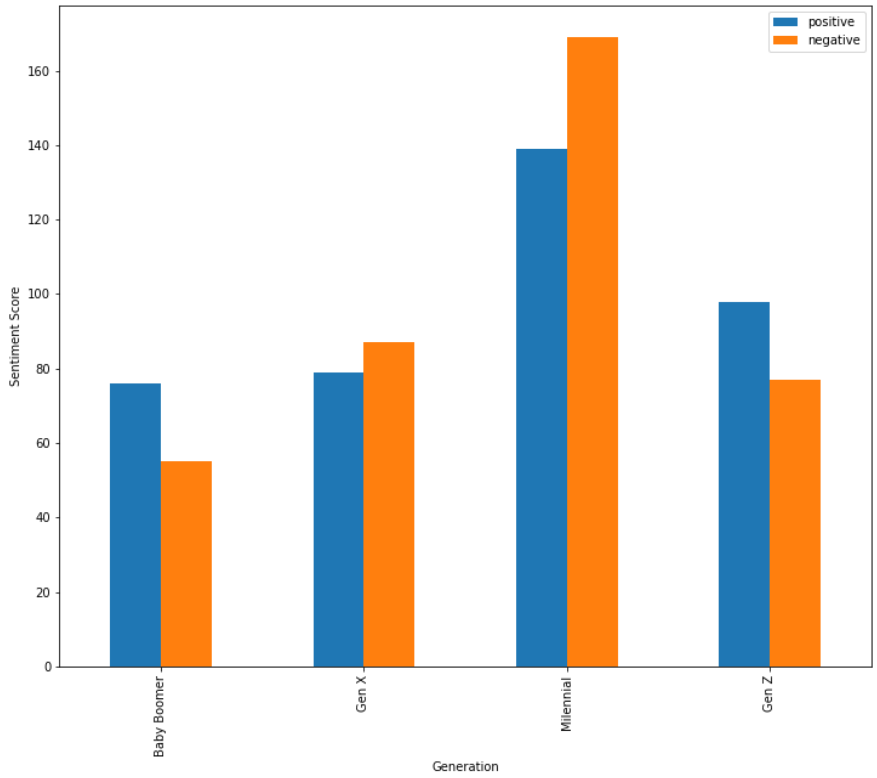
NRC Word-Emotion Association Lexicon for each generation



We see from the results that Baby Boomers show less trust, fear, disgust, and surprise in their emotional content in their lyrics than other generations. We see that Gen X shows a higher proportion of anger and fear in lyrics when compared to Gen Z and Baby Boomers. It's obvious that Millennials show much higher emotional scores across the board, but we see that fear, sadness, and anger are of similar proportion to Gen X. This could be because they share similar artists, namely Nirvana, Linkin Park, Metallica, and Eminem, who are known for darker lyrics.

This theory is further supported by plotting the positive and negative sentiment scores of each generation against each other.

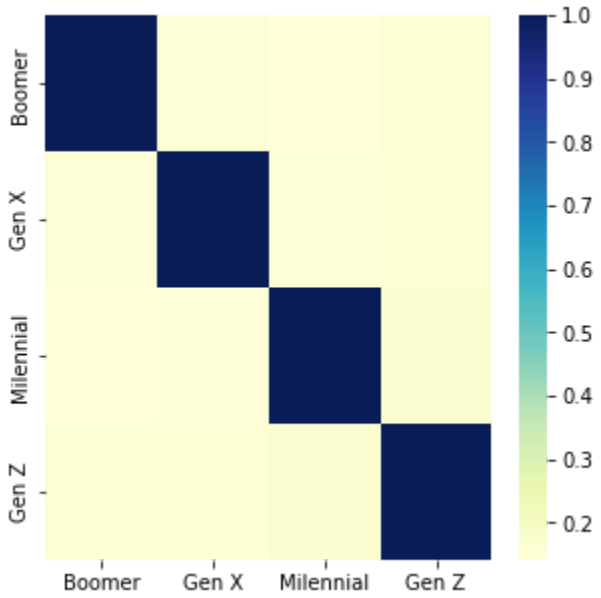
Sentiment Scores by generation



Both Gen X and Millennials showed a higher negative sentiment than positive, where the other two generations had the opposite.

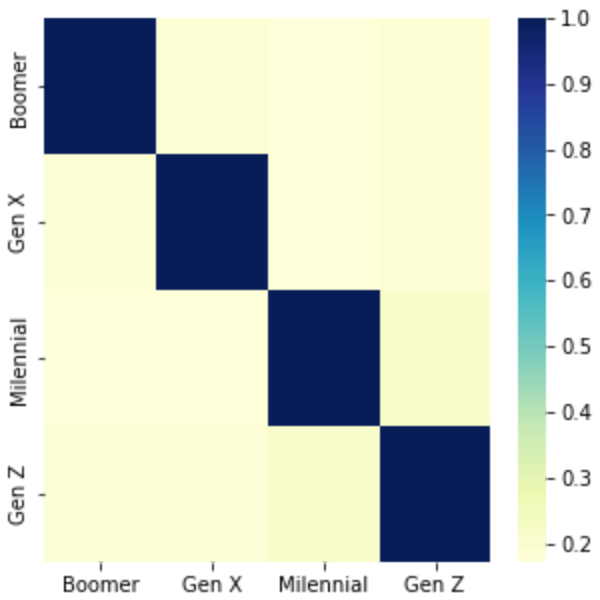
The final analysis to be done is similarity between lyrics. For this, I used sklearn's CountVectorizer, TFIDFVectorizer and cosine similarity functions. Surprisingly, although some of the lyrics between generations show similar emotional sentiment, there is very little cosine similarity between them.

Cosine Similarity between generations (CountVectorizer)



The comparison was done again, but using TF-IDF as a Document-Term Matrix metric this time.

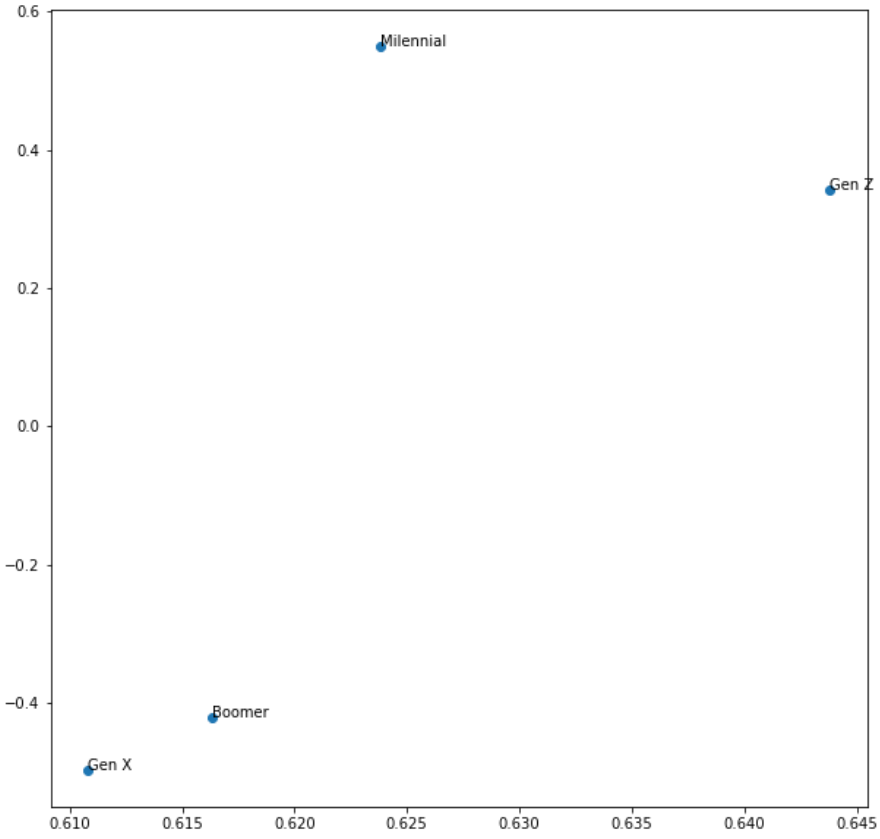
Cosine Similarity between generations (TFIDFVectorizer)



Millennials and Gen Z show very little commonality this time around, but still low overall.

Finally, LDA dimensionality reduction was done on the TF-IDF Vectorizer to plot visually where similarities between generations may lie. From here we see commonalities between Baby Boomers and Gen X.

Generational similarity visualization (TF-IDF)



Discussion:

After thorough analysis of a sample of popular artists for each generation, it can be concluded that they share no significant differences in their lyrical diversity. Millennials did display a higher R value, but this might be attributed to the rapper Eminem, who is known for his lyrical diversity. They also share a difference in emotional sentiment, which is why some people (who also may be more vocal about their musical opinions) may identify more with. The similarities were shared between Baby Boomers and Generation X, and Millennials and Generation Z. These two groupings may be due to their proximities year wise, and the rise of the internet with the latter two generations.

Certain limitations should be considered, especially the lack of a large enough corpus. Gathering more artists (and evidently more songs) would allow our analysis to be more accurate and potentially point out some differences we might not have noticed.

Overall, from the results, it can be concluded that there is no drastic difference between lexical content and implications between the lyrics of the past and the present. Wishing to be born in another generation may then be caused by the successful marketing of the days gone by, where things that seemed normal in the past are used as aesthetic selling points today. A successful nostalgia campaign and the availability of almost every piece of music on streaming platforms can potentially be another phenomenon to be looked at.

Bibliography:

1. https://www.urbandictionary.com/define.php?term=I%20Was%20Born%20In%20The%20Wrong%20Generation
2. https://www.urbandictionary.com/define.php?term=Mumble%20Rap
3. https://chartmasters.org/2020/04/most-successful-artists-by-decade/
4. https://www.azlyrics.com/
5. https://www.crummy.com/software/BeautifulSoup/bs4/doc/
6. https://www.nltk.org/
7. https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/
8. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm