Analyzing Lyrics Through the Decades

Lior Levy Meruk
Mary McSweeney
Henry Johnson

University of California, Los Angeles
STATS 133

August 4, 2025

Abstract

It is a well-known fact that culture and identity play a significant role in the development of musical styles and lyrics (Tran 2023). To test this theory, our team decided to investigate a dataset containing the top 100 songs from the years 1959-2023 in the USA. Our main goal is to investigate and analyze the lyrics of every song and discover trends over time, hopefully related to shifts in our society. To begin, we constructed a corpus from the lyrics column and cleaned the text with various methods. We decided to split the data into decades, in order to make it easier to see patterns, differences, and trends. After performing sentiment analysis, looking at most frequency words and bigrams, using LDA, and even word clouds (all per decade), we then constructed a predictive model to try and predict whether an artist is male or female based on their lyrics alone. Our results show how artists react to the ever changing world around them and how words stay powerful and relevant through all the years.

Contents

Introduction

Musical styles and lyrics have evolved greatly over time. From the emergence of jazz in the 1920s, rock in the 50s, then hip-hop in the 70s, and now to modern pop and rap, different music genres have become more prevalent throughout the decades; and with each distinct genre, comes a unique set of vocabulary that emerges. Not only that, but every decade had its own historical events that shaped our society here in the United States. Through the Cold War, the struggle for civil rights, the moon landing, economic booms and recessions, and many other major events, the United States has experienced vast amounts of change over the past seven decades. With each event, public sentiment and opinions shift, showing a direct reflection and influence on popular music. For example, the civil rights movement in the 1960s showed a profound impact on music, with artists even using their music to call attention to social injustice and promote equality (Tran 2023). Women empowerment is another political and social issue that can be clearly seen through music. More patriotic songs can also indicate times of insecurity and a need for a stronger identity as a nation.

Our group decided to explore these changes and trends through six decades; more specifically, the 1960s-2010s. Our dataset includes the top 100 songs for each decade, with a total of 6500 songs. Each row also details the song artist, album, lyrics, release date and year, and featured artists. Using this information, our team will split the dataset by decade, and then perform further analysis to conclude whether there is a drastic change in lyrics and lyric sentiments in the US during the past six decades (1960s - 2010s). We also have an additional goal of comparing male and female artists, in order to see if their lyrics are significantly different from one another, in an attempt to predict gender based on lyrics.

Data Discussion

Our dataset is the Top 100 Songs & Lyrics by Year in the US from 1959-2024, from the website Kaggle. It contains the top 100 songs in the US from the years 1959 to 2024, according to Billboard Top 100 Magazine. It also includes additional information like the Album, Album URL, Artist, Featured Artists, Lyrics, Media Link (i.e. YouTube link), Rank (at end of year), Release Date (year-month-day), Song Title, Song URL (link of lyrics from genius api), Writers and the Year of ranking. So we have 12 columns in total, and 6500 rows or songs. It is important to note that although our dataset contains information on songs from 1959 to 2024, we are only interested in comparing songs by decade, and the 1950s and 2020s were not complete (incomplete data).

Data Cleaning

As mentioned above, we were only interested in comparing our songs by decade. As a result, our first step was to remove all songs from 1959, as well as 2021, 2022, 2023, and 2024. This gave us a dataset with songs only from our timeframe of interest. Then, we cleaned our data by removing the columns we were not interested in analyzing, and only kept Artist, Lyrics, Song Title and Year. We then added a column exclusively for the song decade. In order to answer our second research question (predicting artist gender from song lyrics), we found another dataset containing the top artists in the past few decades, as well as their gender and category (music genre). This dataset mostly contains modern artists; ones active after the 1990s. However, the artists in this dataset do overlap with the artists in our original dataset. After merging the two, we had 2592 songs and artists classified into a gender. This will enable us to make a predictive model classifying the artist gender from their lyrics.

The Lyrics column of our dataset is our primary column of interest, since it contains the lyrics from every song. We applied the usual data cleaning steps to our lyrics; we removed numbers, converted to lowercase, removed English and custom stopwords, removed punctuation, and we also removed filter words (yeah, ooh, oh, etc.), which were so prevalent in our data and contained little to no meaning. Since we also will be working with the Artist names, we converted the Artist column to lowercase to ensure credibility.

Methods

Word Frequencies

As a preliminary method of comparing the different decades at our disposal, we started by looking at word frequencies. Unsurprisingly, the top words in every decade are "love" and "baby", which are popular lyrics that appear across genres. A big shift is noticeable in the less popular most frequent words, with some unique words emerging, like the use of mam in the 1970s and b*tch in the 2010s (Figure 1). This shift is likely due to more explicit language being normalized and even praised in modern times, and is reflective of important topics like gun violence, misogyny, and social issues being brought up in music. To dive deeper into this hypothesis, we paired up our individual lyrics into bigrams. As seen in Figure 2, exploring popular word pairings proved that indeed, over time popular lyrics prefer more action-oriented verbs (gonna get, wanna see) in favor of the earlier decades, which had more positively-perceived pairings (little girl, go home).

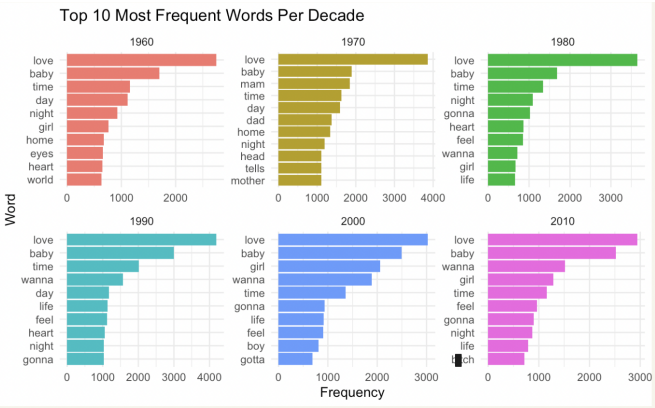

Figure 1: Top 10 Most Frequent Lyrics Words by Decade

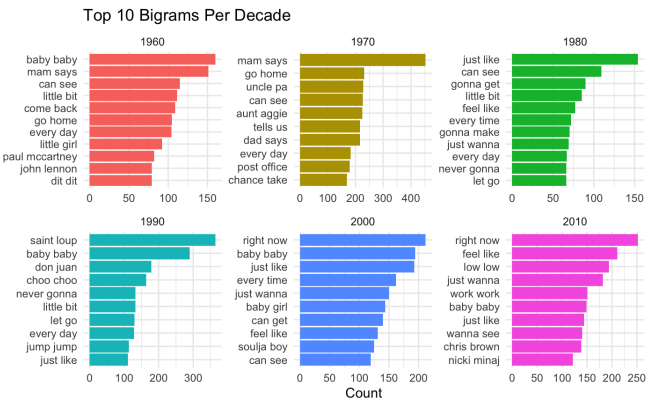

Figure 2: Top 10 Most Frequent Lyrics Bigrams by Decade

Sentiment Analysis

After analyzing word frequencies, our team was curious about how emotions and sentiments have shifted over time. We used the NRC lexicon to extract the top sentiment emotions by decade, and found some interesting results. The top emotion from each decade is as follows: 1960s - Anger, 1970s - Trust, 1980s - Disgust, 1990s - Fear, 2000s - Joy, 2010s - Sadness (Figure 3). Every decade has its own distinct most common emotion, as well as its own less common emotions. This may be due to global or national events, with some decades overall having a less optimistic view of our world and our future. For example, the 2008 housing crisis and proceeding great recession may be partially responsible for the shift from joy to sadness in the 2000s to 2010s.

When looking at Figure 4, which shows only the positive and negative sentiments from the Bing lexicon, another pattern emerges. The 1960s and 70s are stable at around 54% negative sentiments in popular song lyrics. However, the 1980s has a surge in positive sentiments, and the proceeding decades have a steady decline in positive sentiments. From 47.6% overall negative sentiments in the 80s, this increases to 59.2% in the 2000s, with the highest overall negative sentiments present in this decade. As mentioned above, this could be due to a combination of factors; most likely, a mixture of economic, social, and political issues that are reflected in less positive emotions being released through music.

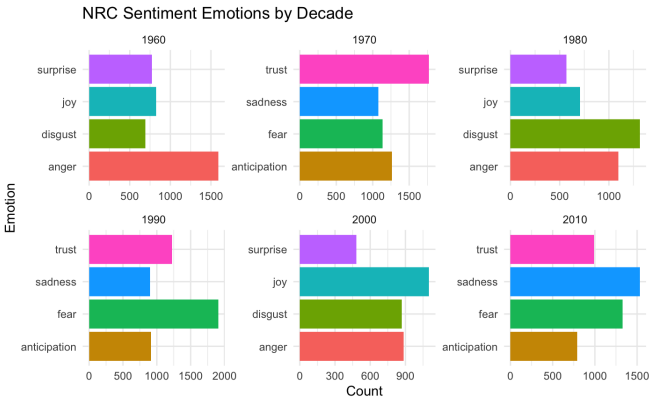

Figure 3: Top NRC Sentiments by Decade

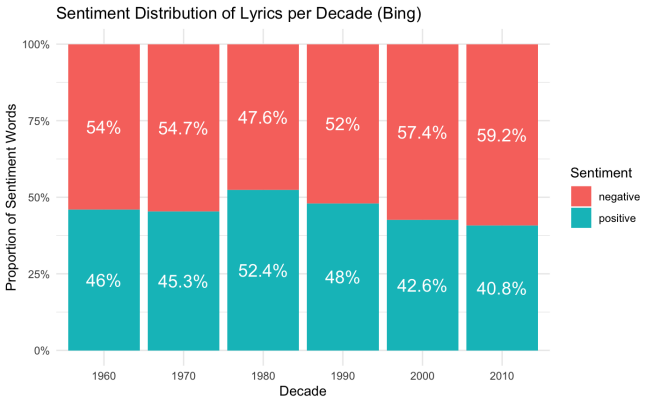

Figure 4: Overall Positive & Negative Bing Sentiment by Decade

Topic Modeling & Classification

LDA

In order to look at how the content of each decade related, we decided to perform Latent Dirichlet Allocation (LDA). First, we created a term document matrix where each song was a document and attached the decade to each song title to keep track of. Then, we ran an LDA with 6 topics to match the 6 decades in our dataset. Looking at the top words for each topic (Figure 5), we can see that the top words do not particularly represent any decade. Instead we found that better labels for our topics than decades would be Hip Hop, Daily Life, Family, Romance, Rap, and Dancing.

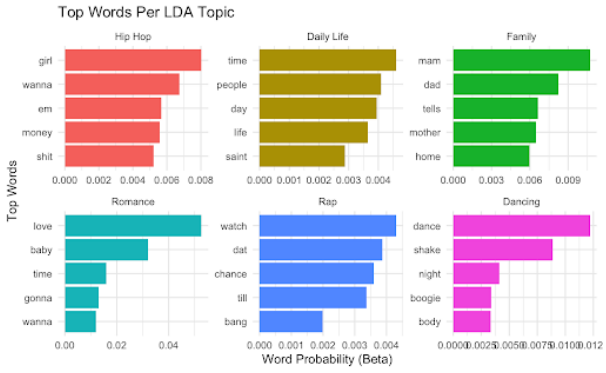

Figure 5: Top Words per LDA Topic

Next, we looked into the distribution of the LDA topics across decades (Figure 6), and found that the Romance topic dominated all decades, with Hip Hop coming in second for every decade as well. Excluding Romance and Hip Hop as topics, we wanted to see if all decades continued to favor the same topics in the same order. Upon closer inspection, we found that all decades had the most songs in Dancing next (Figure 7), but there are some differences after that. The 1960's, 1970's, and 1990's all favored Family, while the 1980s, 2000s and 2010s all favored Daily Life. Rap is the least prevalent topic, however it seems to be the most common (by comparison) in the 2010s.

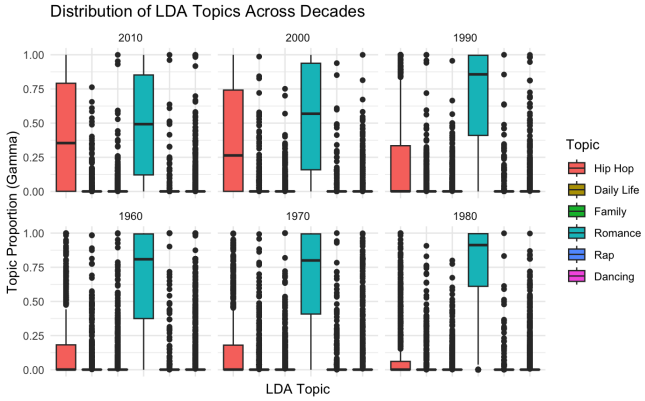

Figure 6: Top Words per LDA Topic

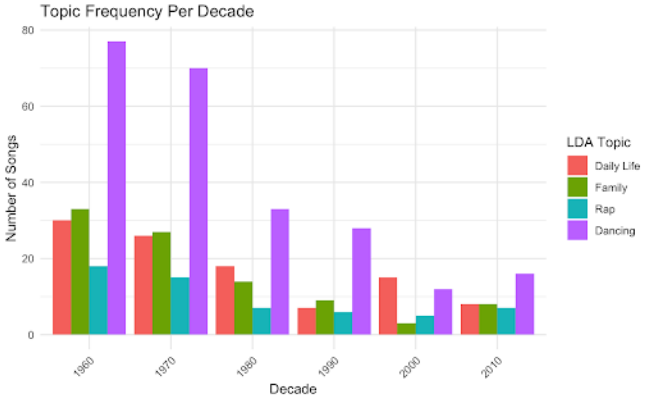

Figure 7: Top LDA Topic Frequency by Decade

We classified songs into topics by looking at their gamma values and choosing the most likely one. Then, we classified decades into topics by choosing the topic that appeared for the most songs that decade. We then found the number of misclassified songs by filtering to rows where the topic classification for the song does not match the topic classification for the decade. Since our findings from LDA were that the each decade was in topic 4, Romance, any song classified in any other topic was considered misclassified. This gave us a misclassification rate of 83% and classification accuracy of 17%. This means that around a sixth of our data was correctly classified, and as we had 6 decades, this is about equal to random guessing and our model was not able to distinguish between decades.

Classification Model

Since our attempt to split our data into different topics by data was unsuccessful, our team was curious to see if we would have better luck classifying songs by gender. We joined the Singer's Gender dataset from Kaggle to our original dataset in order to apply a Random Forest Model. We chose to do a 70% training and 30% testing split to train our Random Forest Model to predict gender based on lyrics.

We can see through our confusion matrix that our model did a good job classifying males, with 439 correctly classified and 40 misclassified, while it had more trouble with classifying females, with 116 correctly classified and 183 misclassified. This could be because our dataset had more data on males than females. To further visualize our confusion matrix, we have included a heat map where a darker shade indicates a greater number. This shows that the model primarily predicted male and was correct the majority of the time. We achieved an accuracy rate of 71.34%, which is greater than just random chance, meaning that male and female artists differ in lyrical style and are classifiable.

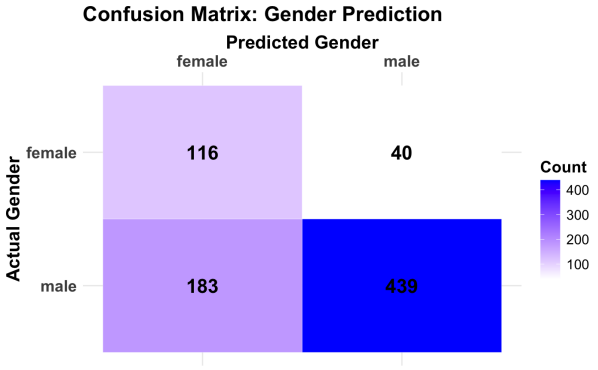

Figure 8: Gender Prediction Heatmap

Conclusion and Limitations

Conclusion

By analyzing song lyrics from the 1960s to the 2010s we were able to see surprising trends and cultural shifts that reflect societal changes during these times. We see words like "love", "baby", "time", and "girl" stay in use throughout all decades reflecting the power and importance of these words and topics to us. In our sentiment analysis, we see our sentiment emotions transitioning from mostly happy and joyful (1980s) to more sad and negative emotions in the 2010s. Our LDA revealed that lyrics are not inherently grouped by decade, but perhaps by other themes such as song genre or lyrical topic. Our fairly successful gender-based predictive model shows how gender can influence song lyrics and the differences between male and female artists.

The greater significance of this project is seen when additional research is brought in. As music evolves, so does the message that is being sent and extracted from songs. As seen from our sentiment analysis and most frequent words per decade, more recent lyrics favor curse words, actions, and inappropriate lyrics. "With heavy, catchy beats, artists can sneak by with inserting inappropriate language in their lyrics" (Hirashima 2023). With censorship being a prominent issue, perhaps artists are fighting back against those limitations, resulting in more explicit messages and words being used. However, the impact can be felt, especially among younger listeners that are subjected to explicit imagery and messages in popular songs (Hirashima 2023). In recent decades, not only are we subjected to those lyrics, but overall sentiment is more negative and less optimistic.

Limitations

Our dataset contains the Top 100 Billboard songs for each year. So intuitively, our data may not fully capture all kinds of music styles and underground music. If we wanted to analyze each decade's music as a whole, this dataset may be too limited, and we would want to find a dataset with a wider range of songs. Our data only covers songs from the 1960s to the 2010s, so it is possible that if we had more data we would be able to see more complete and obvious trends. Since our sentiment analysis is based on predefined lexicons, our analysis may fail to capture the evolving meaning of words or how a word can change meaning in context, like modern slang for example. There are certain factors our project does not cover like industry trends, production styles, or target audience information which may all affect our analysis and conclusions. For further analysis, our group may consider using a more diverse dataset, and we may also explore the different popular genres of each decade and how they differ. Another potential analysis could be to explore the difference between male and female artists (sentiment analysis, LDA), and how that difference has changed through the past few decades.

Bibliography

Blakely, Brian. (2021). *Top 100 songs and lyrics from 1959 to 2019*. Kaggle. `https://www.kaggle.com/datasets/brianblakely/top-100-songs-and-lyrics-from-1959-to-2019`

Weinberg, John. (n.d.). *The Great Recession and its aftermath*. Federal Reserve History. https://www.federalreservehistory.org/essays/great-recession-and-its-aftermath.

Hirashima, Phoebe. (2024, March 1). *Beauty of lyricism lost in the heat of trends: Effect of explicit music on listeners*. https://imuaonline.org/2480/editorials/beauty-of-lyricism-lost-in-the-heat-of-trends-effect-of-explicit-music-on-listeners/.

Kibria, Rashedul. (2022). *Singers Gender*. Kaggle. `https://www.kaggle.com/datasets/rkibria/singersgender`

Tran, Kristine. (2023, May 3). *How Music is a Reflection of Society*. Melody Studio. https://melodystudio.net/2023/05/03/how-music-is-a-reflection-of-society/.