Mapping Cultural Evolution Through Song Lyrics

Nicole Kloss, Jan Thiele, Eszter Diána Kocsis, Katrin Sauter

1. Introduction

Our project, "Mapping Cultural Evolution Through Song Lyrics," investigates the changing themes in popular music, with a specific focus on the period surrounding the COVID-19 pandemic. Music is a central part of everyday life, and popular genres often reflect the social, cultural, and economic contexts in which they were written. The global impact of the COVID-19 pandemic on both the economy and all levels of society makes it an ideal case study for using large language models to uncover meaningful cultural trends and connections.

We followed an exploratory research approach to address broad research questions, such as: "*How do lyrical themes evolve over time?*" and "*How do they relate to social and economic contexts?*"

For our methodological approach, we used sentence embeddings to analyze lyrical content, a method that captures the semantics and context of lyrics, which is especially useful for the metaphorical and figurative language found in songs. This approach has significant advantages over traditional statistical methods used for uncovering themes in textual data (e.g. word clouds or topic modelling). Our use of embeddings builds on the work of Pizzaro et al. (2024), who used BERT for genre classification and success prediction, and McVicar et al. (2021), who used embeddings for music tagging, genre assignment, and explicit content flagging.

Our primary contribution is the development of a nuanced and structured embedding space, which allows for detailed semantic analysis of lyrics, eliminating the need for manual annotation or external tagging. This allows for a more fine-grained examination of how lyrical themes relate to broader socio-cultural factors, while still ensuring our model can effectively classify songs by genre. In the following section, we describe how we collected and prepared our dataset, which serves as the foundation for the analyses presented in this study.

2. Data Collection and Preprocessing

2.1 Data Sources and Collection

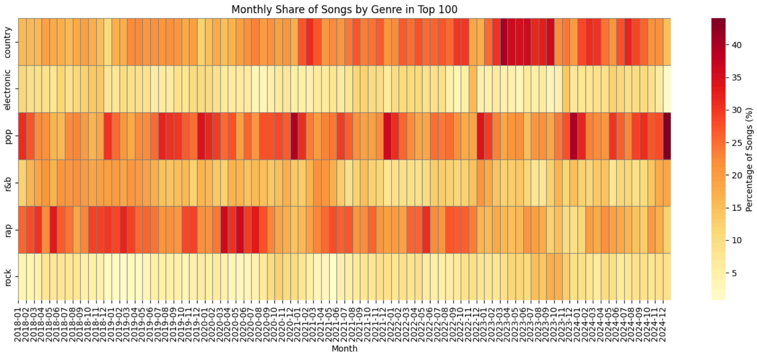
We collected data from two sources: Billboard Charts' weekly Top 100 list and the Genius API (Genius, n.d.) (where we retrieved the songs based on the Billboard listings). From the Billboard charts, we collected data starting from the first week of January 2000 up until early August 2025. For this, we used BeautifulSoup for webscraping from the Billboard webpage, which took roughly an hour due to the large time span. The collected dataset initially contained 133,600 entries, which was reduced to 11,549 unique songs after removing duplicates. In order to use the Billboard data for collecting lyrics from the Genius API, we had to normalize the artist names, since Genius is very strict about exact name matches. For this, we tried to build a comprehensive regex function that tokenized the connectors used in collaborations (e.g., 'and', '&', ',', or 'featuring'). This worked fairly well overall, except for cases with an 'X' between names. In those situations, it was almost impossible to tell whether the 'X' was part of the artist's name or meant to indicate a collaboration. Similarly, when analyzing missing lyrics, we found that tokenizing ‘and’ caused issues — for example, Ariana Grande (“*Gr and e”*) songs were not scrapped at all. These cases had to be corrected manually later on. After that, we ran the API calls for the songs, which, due to the large amount of data, took approximately 12–14 hours to complete. From Genius, we retrieved complementary data such as the song’s language, genre (primary_tag), release date, album, and lyrics. After completing the data collection and removing songs with missing lyrics, 9,490 remained out of 11,459 unique songs.

2.2 Data Cleaning

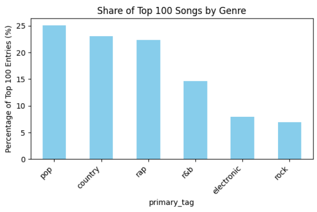
After data collection, the dataset was restricted to the period from 2018 to 2024, covering the years preceding, during, and following the COVID-19 pandemic in order to capture potential shifts in lyrical trends associated. We merged the two datasets, and by using the unique songs from the Billboard data, the Genius dataset was filtered to retain only relevant entries. Non-English songs and non-music entries were excluded, while all genre tags were normalized to lowercase for consistency. We cleaned the lyrics in two main steps. First, using regex, we removed editor comments/notes appearing at the beginning of the lyrics. We also filtered out markers such as [Intro], [Verse], or [Chorus], as well as any text in parentheses (and parentheses themselves too), which usually contained repetitions, interjections, or stopwords. Extra whitespaces were also removed. Once the lyrics' structure seemed sufficiently cleaned, the second step was to remove English stopwords. We extended this stopword list with additional meaningless words commonly found in lyrics, such as ‘oh’, ‘yeah’, ‘na’, ‘woo’, etc. Only one song had to be removed from the list due to missing lyrics. Finally, based on the Billboard columns' *rank* and *peak,* a new column, *peak_date,* was created to record the week when each song reached its highest chart position. After cleaning, the final dataset consisted of 3,195 songs.

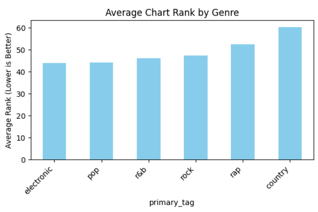
2.3 Exploratory Analysis

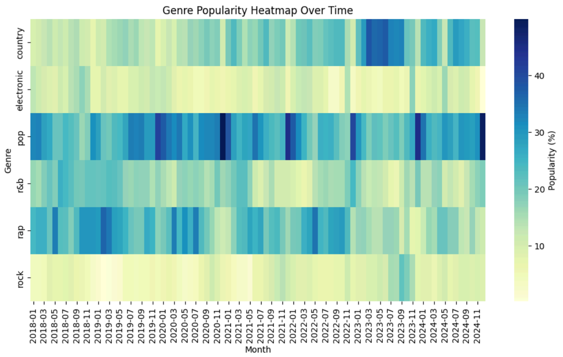
To gain an initial understanding of the lyrical and cultural patterns within our dataset, we conducted an exploratory analysis focusing on genre distributions, theme identification, and temporal trends. For the exploratory analysis, a new dataset was constructed comprising all weeks and their respective top 100 songs, to which the same data-cleaning steps described above were applied. 26,636 songs from 1094 artists remained, averaging 3,805 songs per year, with a roughly equal yearly distribution ensuring a balanced dataset for analysis. The annual average is lower than the theoretical maximum of 5,200 (100 top songs for each of the 52 weeks), likely due to the removal of non-English songs, exclusion of rare genres, and filtering out non-music entries from the Genius API. Genre-wise, pop dominates the dataset, followed by country and rap, with R&B, electronic, and rock trailing behind. Examining the relative share of each genre among the top 100 songs reveals that pop, rap, and country hold the largest portions. Pop and rap alternated as the most prevalent genres until 2021, when country music gained prominence. Country music showed the largest growth over the years, with a particularly strong year in 2023, while rap experienced a decline in popularity, especially after 2022. It could be further observed that pop songs tend to peak toward the end of each year, potentially reflecting the seasonal popularity of Christmas-themed music. R&B share declined until 2022 before stabilizing, whereas electronic and rock consistently maintained smaller shares, with rock gaining



some popularity in the latter half of 2023. The average song comprises 181.5 words, with rap songs being the longest at 264 words, while the other genres show comparable lengths. A slight downward trend can be observed over time, along with seasonal fluctuations, indicating that songs tend to shorten toward the end of the year and lengthen at the start of the next year. 2020 and 2021 show slightly higher averages, influenced by longer R&B tracks in 2020 and rap songs in 2021. Further analysis of genre popularity shows that songs reaching rank 1 are dominated by pop (38%), followed by rap (17%), country (14%), R&B (13.5%), rock (11%), and electronic (6%). This aligns with the observation that pop also dominates in terms of Top 100 chart representation, followed by country and rap. Interestingly, electronic songs, while relatively rare, achieve the highest average ranks, suggesting that the few songs that reach the top 100 often perform exceptionally well. To provide a more robust comparison, a weighted popularity measure was computed by assigning ranks descending weights (100 points for rank 1 down to 1 point for rank 100), summing them across weeks, and normalizing into percentages. This measure confirms pop as the most successful genre overall, followed by rap and country. Over time, country music has been gaining traction, culminating in a particularly strong year in 2023, while rap and R&B show signs of decline in the examined period. Pop exhibits prominent seasonal spikes in December, likely driven by Christmas-themed songs.

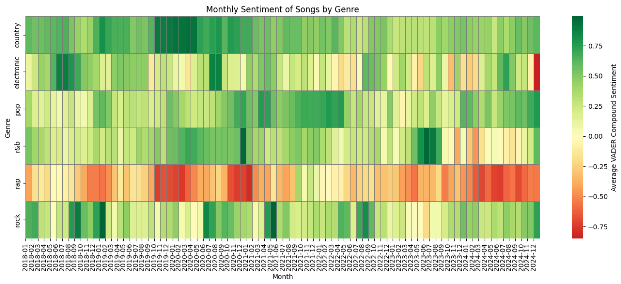


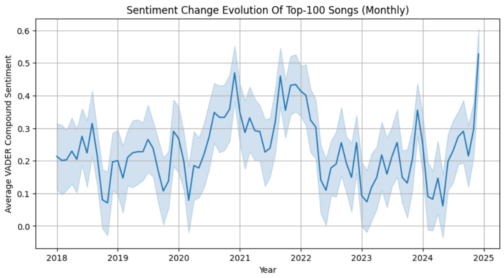




An analysis of song persistence in the Top 100 shows that, on average, songs remain for 8.4 weeks, though the vast majority only appear once. The most enduring hits can accumulate up to 91 weeks in the charts. This measure though does not imply continuous persistence and some may be explained by seasonal re-entries (e.g., Christmas music). At the genre level, country songs prove the most persistent, while rap songs tend to drop out more quickly and electronic as well as rock songs display very ambivalent persistence patterns.

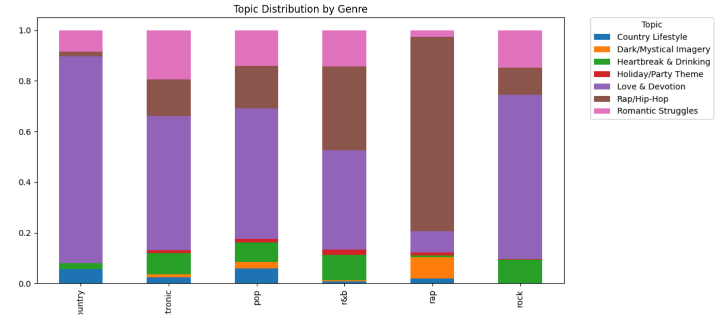
Sentiment analysis using VADER (focusing on the compound score: −1 = most negative, +1 = most positive) reveals rap as the most negative genre and country as the most positive, followed by pop. No clear long-term trend can be observed across years. However, 2021 stands out as the most positive year with the highest average compound score (0.33), followed by 2020 (0.27) and 2022 (0.25). The temporal evolution of sentiment reveals a sharp increase in 2020, followed by a decline in early 2021, a brief recovery, and another drop in early 2022. However, analyzing sentiment trends by genre indicates that these fluctuations cannot be directly attributed to sentiment changes within a single genre. Recurring dips at the start of each year likely reflect the disappearance of positively rated Christmas songs from the charts.

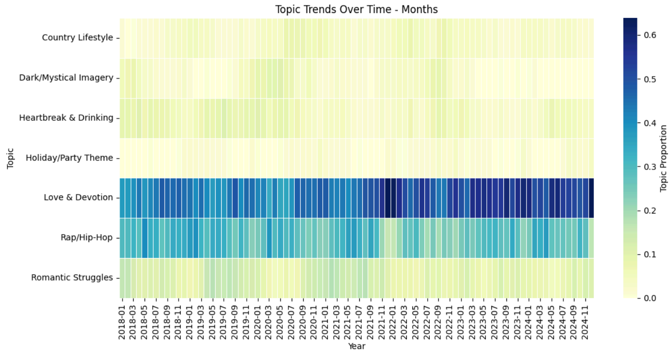




To gain an initial understanding of vocabulary and themes across different genres, word clouds were generated for each genre. This approach highlights the most frequently occurring words, providing a high-level overview of the language patterns used within the genres over time. For improved interpretability, stopwords were tailored to the vocabulary of popular music and removed during preprocessing, allowing the analysis to emphasize meaningful content-specific words. The most frequent words are summarized in the table below. In pop music, the top words—*love, time, baby, feel, back*—reflect themes of romance and personal emotions, consistent with the genre's overall focus on relationships and everyday experiences. Country shows a similar pattern, also indicating storytelling around love and personal narratives (*love, time, girl, back, night, baby*). Rap lyrics feature words such as *nigga, bitch, shit, fuck*, reflecting the genre's use of expressive, often confrontational or street-oriented language. In R&B, the prominence of *baby, love, feel, time, nigga* underscores themes of affection, intimacy, and identity, while Rock lyrics emphasize words such as *back, love, god, away, thunder*, suggesting both personal reflection and broader, sometimes spiritual themes. Electronic music focuses on words like *woman, love, baby, night, dance*, which can be closely linked to nightlife, dancing, and social interaction. Analysing wordclouds over time (different years) reveals that the thematic composition of genres remains largely stable. Occasional spikes in specific words reflect particularly popular songs, such as *watermelon* for pop in 2020 (*Watermelon Sugar* by Harry Styles). Rock and electronic lyrics display the biggest variability across years, likely due to fewer songs in the dataset and the therefore outsized influence of popular tracks. While word clouds provide an intuitive snapshot of word frequency, they fall short in capturing how words cluster into broader thematic structures. To address this, Latent Dirichlet Allocation (LDA) was applied to uncover latent topics within the lyrics, through identifying groups of words that frequently co-occur. Before applying LDA, the lyrics were tokenized and stopwords were removed, with additional customization to exclude common but non-informative lyrical fillers such as "woah." After experimentation, seven distinct themes emerged. The most dominant theme across the dataset is *Love & Devotion*, which prevails in all genres except rap, followed by *Romantic Struggles*. Rap, in contrast, displays a unique thematic profile characterized by *expressive and often aggressive language*, reflecting themes of aggression, sexuality, and street culture, frequently conveyed through explicit content. Interestingly, this theme is also present in other genres, though in a less pronounced way, while it is almost absent in country music.

Topic Interpretation	Top Words
Rap / aggressive style lyrics	bitch, nigga, fuck, money, gang, play
Love & Devotion	love, night, girl, heart, god, christmas, sweet
Romantic struggles/ Heartbreak	love, lie, alone, miss, sorry, club
Breakup/ Heartbreak/ Alcohol Consumption	babe, lose, dance, whiskey, pour, heart
Country lifestyle	taste, jack, bottle, truck, kissin, morgan
Holiday (Christmas) & Party	bell, jingle, sleigh, airport, party, alcohol
Dark and mystical themes	smoke, ghost, wicked, holy, supernova, guitar





When analyzing topic trends across the full dataset, no systematic temporal patterns emerged. The analysis did, however, confirm *Love and Devotion* as the most prominent theme overall, followed by *Rap / Expressive-aggressive Language* (particularly in rap) and *Romantic Struggles*. A more differentiated picture appeared when examining genres separately. In R&B, the thematic focus alternates primarily between Love and Devotion and Rap/ Expressive-aggressive Language. Rock is similarly dominated by Love and Devotion, while expressive-aggressive topics show a slight downward trend over the period considered. Here distinct phases can be observed, including a focus on heartbreak and drinking between mid-2019 and mid-2020, followed by a stronger emphasis on heartbreak and struggles in 2021. Electronic music shows the highest variability in topic distribution. Love and Devotion dominates until mid-2019, after which Romantic Struggles take precedence until mid-2021. This is followed by a return of Love and Devotion until mid-2022, before shorter phases of Expressive-aggressive Language, Country Lifestyle, and Dark Themes emerge. From summer 2023 onwards, Love and Devotion again becomes the central theme. Pop shows a relatively stable distribution with Love and Devotion as the dominant theme, though a small seasonal spike in holiday-related songs can be observed at the end of each year. Country music also centers overwhelmingly on Love and Devotion, with only minor thematic variation. Rap consistently revolves around Expressive-aggressive Language, reflecting the genre's stylistic and cultural specificity. It should be noted, however, that the most variable dynamics appear in genres with smaller sample sizes, where the influence of individual songs is greater. This likely limits the generalizability of the observed trends.

The exploratory analysis provided valuable first insights into the distribution of genres, persistence of songs, sentiment trajectories, and thematic structures across the Top 100 charts from 2018 to 2024. It highlighted the dominance of pop, the rising prominence of country, and the gradual decline of rap and R&B, while also confirming the centrality of *Love & Devotion* as the prevailing lyrical theme across most genres. At the same time, the analysis suffers from the limitations of word clouds and LDA. Word clouds rely purely on word frequency and require substantial domain knowledge to be interpreted meaningfully (for example terms like "watermelon" or "n**ga" in rap culture). LDA, while more structured, depends on co-occurrence and produces topics that remain highly subjective, as themes must be manually extracted and annotated. Therefore, considering these shortcomings, word clouds and LDA are limited in uncovering the deeper thematic meaning within metaphorical and figurative data such as song lyrics. To address this, we decided to use a Sentence transformer model (SBERT) to create a more sophisticated representation of our data. Embeddings of encoder only models are well suited for this since they capture context and semantic similarities rather than just frequency and co-occurrence of words (De Marzo, 2025, p. 13). Lastly, embeddings provide a nuanced view of lyrical content, allowing us to track how themes evolve over time and relate to social and cultural contexts.

3. Deep Learning Model

3.1 Training Objective:

Our training objective is to finetune a deep learning model on the themes of song lyrics to produce semantically rich embeddings using contrastive learning (De Marzo, 2025, p. 21). Our goal is to create an embedding space where songs of similar lyrical content are closer than those of different themes. We used the sentence-transformer library and the SBERT model 'all-miniLM-V6-v2' (Reimers & Gurevych, 2019). We decided to finetune the model with genre affiliation as a guiding label. Using genre as an approximation for lyrical themes is not entirely accurate. Genre typically describes musical style rather than contents of a song, which could introduce noise into our data, when pulling together songs of the same genre closer together even though their lyrical themes don't match. However our EDA showed dominant terms within each genre justifying the assumption that genre can serve as a proxy for theme. We could have used the topics generated during the LDA analysis. We decided against this approach, because while nuanced and spanning multiple genres, LDA topics are highly subjective. These topics do not represent a reliable ground truth we can evaluate and test our model against. Additionally, SBERT not only optimizes for the label given but also for semantic similarity. Therefore we expect genre clusters but also overlapping regions where songs of different genres but similar content reside.

3.2 Preprocessing:

We used an extended dataset containing lyrics from 2000 to 2025 (N=9490). The data was split into 70% train and 15% validation and test set resulting in n=6643, n=1423 and n=1424 respectively. We stratified the splits by genre resulting. The data shows a strong class imbalance with the majority label being almost five times as large as the minority one. We tried to address this imbalance by up- and downsampling. Upsampling the data with our chosen loss function creates repetitive training signals which do not provide additional learning opportunities for the model. Downsampling the data resulted in desired training behavior. Training and validation loss curves decreased and cosine similarity accuracy increased. However the resulting performance was worse than with the full dataset. Therefore we decided against these approaches and kept the full imbalanced dataset. Further preprocessing included constructing a triplet dataset of the validation split. We matched each lyric with one of the same genre and one of a different genre, resulting in an anchor, positive, negative structure.

3.3 Loss Function and Training Arguments:

The sentence transformer library provides many loss functions to implement contrastive learning. For our purposes we require a loss function which pulls songs of similar theme and genre closer together while pushing the opposite away creating a nuanced structured embedding space. We compared different loss functions and selected Triplet Loss as appropriate for our goal because it considers the relative distance of a sample to similar and dissimilar samples. Triplet Loss (Reimers & Gurevych, 2019) calculates the distance between a sample (anchor), a similar sample (positive) and dissimilar one (negative). It tries to minimize distance between positive sample and the anchor while maximizing the distance to the negative sample plus a predefined margin. Every negative sample within the margin contributes to the loss and provides a valuable training signal. Triplet loss is unfortunately limited by the quality and diversity of triplets provided for training. We used BatchAllSemiHardTripletLoss (Reimers & Gurevych, 2019), which uses a sample and label structure as input and constructs all valid triplets during training. This alleviates the need for manual triplet construction while also providing the model with a much more diverse training signal. Semi-hard refers to

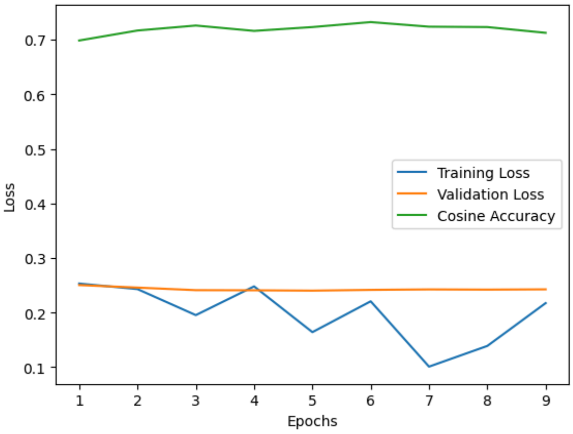
Label	Training Set	Validation Set	Testing Set
0 = pop	1723	369	369
1 = r&b	942	202	202
2 = rap	1724	369	370
3 = electronic	360	77	77
4 = rock	777	167	166
5 = country	1117	239	240

negative examples which are within the margin but are still farther away from the anchor than the positive example. This way the model learns from balanced examples which are neither too hard nor too easy.

We trained the model for 10 epochs, with a batch size of 64 and learning rate of 2e-5. All hyperparameters were fine tuned by manually testing out. Training and validation loss were tracked alongside training as well as accuracy based on cosine similarity. Cosine similarity accuracy measures how often the model correctly assigns higher similarity scores to positive (anchor–positive) pairs than to negative (anchor–negative) pairs (Reimers & Gurevych, 2019). We used the Tripletevaluator native to the sentence transformer library, which evaluates the models performance on manual constructed triplets of the validation set (Reimers & Gurevych, 2019). Using accuracy based on cosine similarity over validation loss gave a more precise measurement for finetuning process, because we evaluated the models ability to classify songs into genres. Additionally we implemented an Early Stopper with a patience of 3 epochs and threshold of 0.001 on cosine similarity accuracy. We used a random sampler rather than the recommended grouped by label sampler. Earlier attempts using the recommended sampler lead to homogenous samples and a training loss of 0.

3.4 Training:

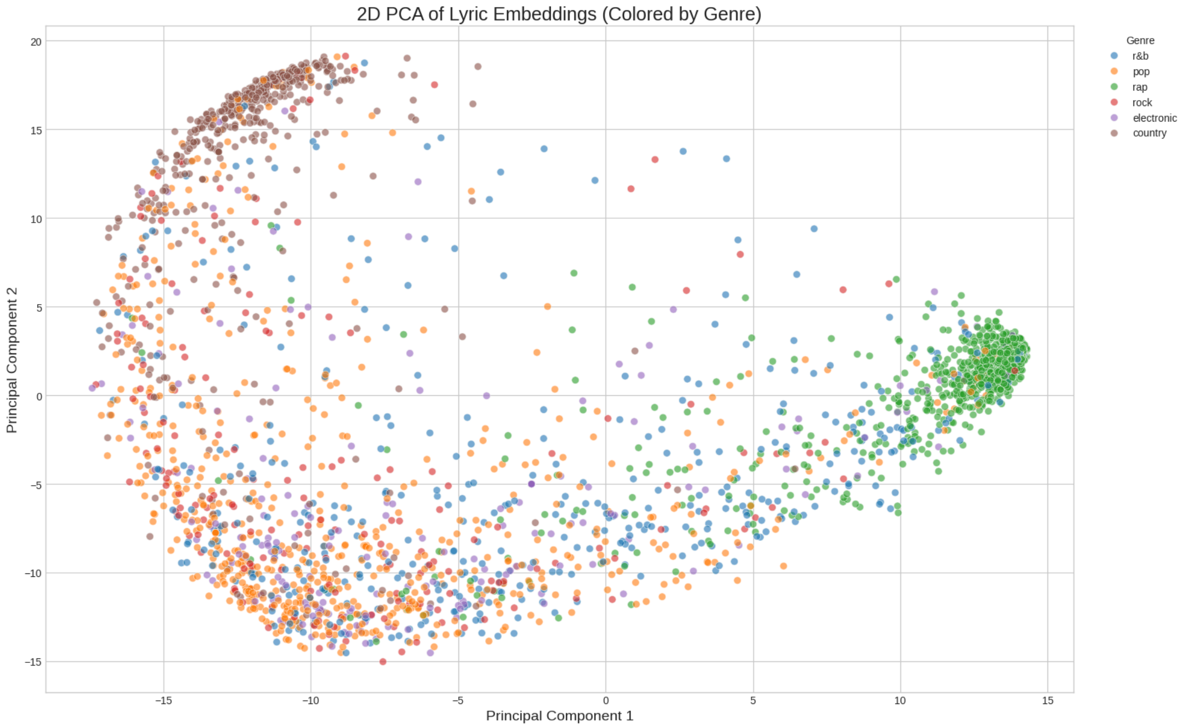
The plot shows the training and validation loss curves as well as the cosine accuracy evaluation during training. Cosine accuracy increases slightly over the first few epochs, reaching its peak around the sixth epoch. Most performance gains were made during the first and second epoch. Training loss fluctuates but overall decreases indicating that the model learned the patterns present in the training data. This instability is likely due to our sampling technique. The random sampler can construct batches of different difficulty resulting in the fluctuating loss seen in the plot. On the other hand validation loss did not decrease and plateaued. This could indicate overfitting because of training curves decreasing simultaneously. However this seems unlikely since the cosine accuracy on the validation set still improved over the course of training. The validation loss plateauing could be explained by inspecting the construction of our validation sets: it was smaller and less diverse than the training set providing less training signal overall and more sensitive to class imbalance. Moreover cosine accuracy was calculated on a manually constructed triplet version of the validation set which reduced diversity which may have biased our evaluation scores towards higher measurements not reflected in validation loss.



3.5 Testing and Validation:

We compared the base and the finetuned model on two different metrics for the validation and test set. Similar to training we compared the cosine similarity accuracy for both models. The finetuned model showed better performance across both datasets, indicating that the finetuned model was much more likely to place positive samples closer to the anchor than negative samples compared to the base model. Additionally, we used the encoded lyrics of our validation and test set to train a simple K-nearest neighbour classifier with genre as a target. We hypothesized that the finetuned model produces embeddings that are better suited for genre classification since we used genre as a guiding variable during training. Again, the finetuned model outperformed the base model. These tests confirm our fine tuning approach for the genre was successful.

Model	Accuracy Cosine Similarity		KNN Accuracy (N=5)	
	Validation	Test	Validation	Test
Base	0.6016	0.5934	0.49	0.49
Fine-Tuned	0.7323	0.7051	0.57	0.56

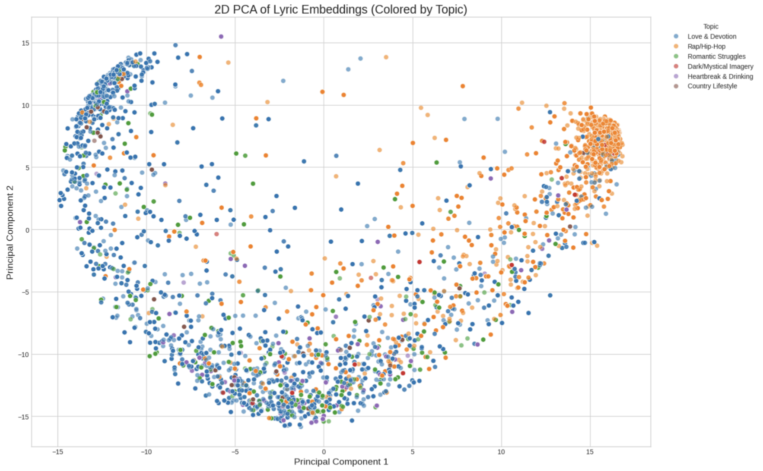


Beyond genre classification we evaluated the models embeddings in a qualitative way. After encoding the song lyrics with the finetuned model, we reduced their embeddings to two dimensions using Principal Component Analysis and plotted the resulting components. The plot shows the rap genre thematically distinct from other genres along the axis of the first principal component, with rap lyrics (green) forming a distinct cluster.. Country and Pop appear distinct along the second principal component while Rock (red), R&B (blue) and Electronic (brown) music overlap, suggesting they share common lyrical themes aligning with our EDA. These results align with our fine tuning goal: Embeddings show genre affiliation but they also preserve thematic similarity across genres. This means neither genre nor semantic similarity dominate the embedding space. The model captures both which validates our finetuning approach beyond our quantitative measures before. Therefore we created a much more sophisticated representation of song lyrics suitable for more detailed analysis beyond our initial EDA.

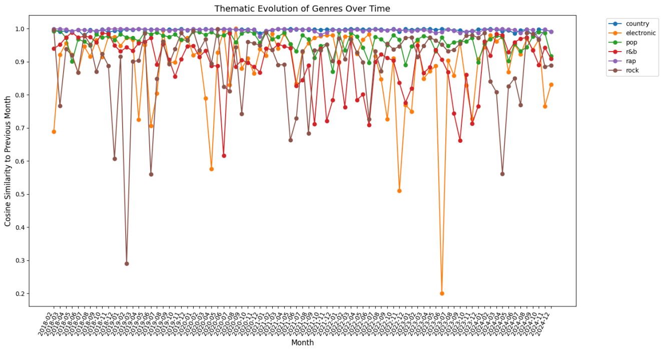
4. Cultural Studies Analysis

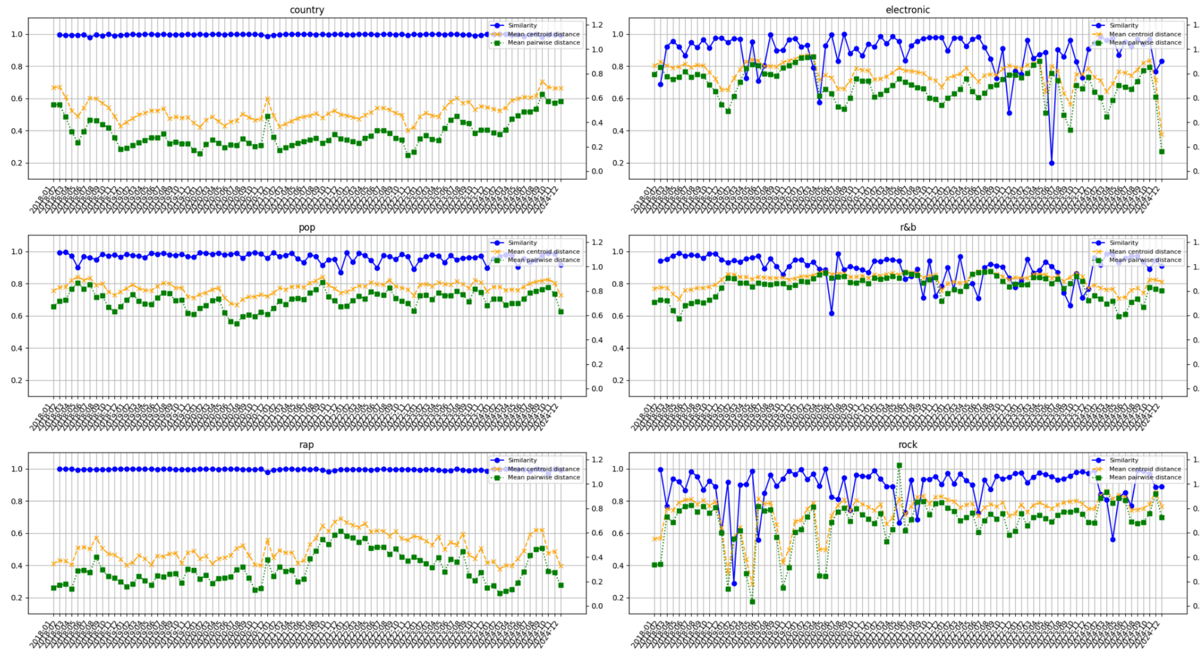
4.1 Tracing Thematic Evolution in Popular Music Through Lyric Embeddings

To investigate the evolution of musical themes over time, we created a "thematic map" using the reduced song embeddings from before. However in this plot, the dots are colored according to the topics defined in the exploratory analysis. *Rap/Hip-Hop* lyrics (orange dots) form a dense and distinct cluster on the far right, suggesting a distinct lyrical style from the other topics. This not only aligns with the insights from the exploratory analysis but also with the PCA plot above, where dots are colored according to genre, in which the rap genre largely overlaps with this same cluster. By contrast, *Love & Devotion* lyrics (blue dots) spread broadly across the left side, indicating a wide thematic range with diverse lyrical expressions. Related themes such as *Romantic Struggles*, *Heartbreak & Drinking*, and *Dark/Mystical Imagery* are interspersed within the *Love & Devotion* area, suggesting a thematic overlap. For instance, heartbreak and romantic struggles are often related to the theme of love.



To extend this initial analysis, embeddings were further analyzed to track their trajectories across genres and artists. Distances between embeddings were quantified using cosine similarity, which measures the angle between vectors in high-dimensional space and is insensitive to differences in magnitude. This makes it well-suited for comparing semantic content, as it captures thematic similarity regardless of the length or intensity of individual embeddings, making it possible to track shifts in thematic content, identify periods of experimentation and assess stylistic consistency. At the genre level, thematic stability over time was analyzed by computing the mean embedding per month-genre pair and comparing it to the preceding month. The results indicate that country and rap exhibit the highest stability, followed by pop, whereas R&B, rock, and electronic show greater variation in thematic content. However there are no visible trends emerging for any of the genres during the observed time period. To further quantify diversity within genres, for each month-genre pair the mean centroid distance, reflecting the spread of songs around the average theme, and the mean pairwise distance, representing the average dissimilarity between all song pairs, were computed. Within each genre, both measures followed similar trajectories over time, suggesting that the observed patterns of thematic diversity are robust and not dependent on the choice of metric.



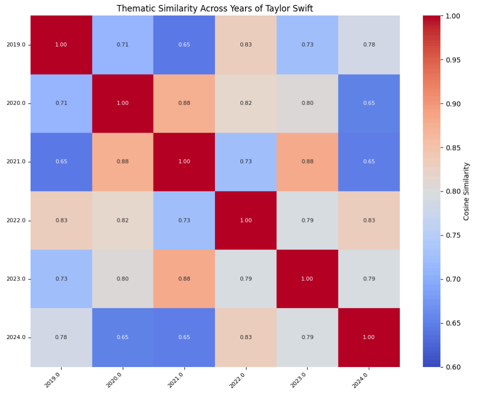


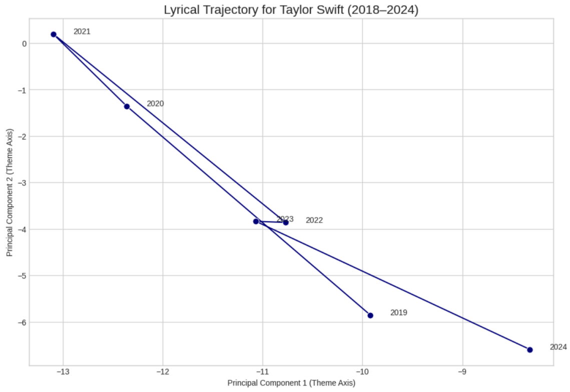
Country and rap consistently displayed the lowest diversity, while R&B and pop exhibited the highest, and electronic and rock showed the greatest fluctuations in thematic diversity. Rock experienced substantial variability from late 2018 to mid-2020 before stabilizing, suggesting a period of experimentation or shifting sub-genres. Rap showed an increase in diversity during 2022, which remained elevated until a dip in fall 2023, while R&B experienced a decrease in diversity in 2024. Country demonstrated a gradual increase in diversity beginning in 2023, coinciding with a rise in the number of songs entering the top charts, which may have contributed to higher measured diversity. In contrast, for genres such as rap and R&B, which saw an overall decline in chart popularity, a corresponding decrease in diversity was not observed. This indicates that changes in popularity do not necessarily mirror thematic diversity within a genre, and fluctuations in diversity may reflect intrinsic stylistic variation rather than solely popularity-driven effects. It has to be noted however, that this analysis is limited to songs that reached the Top 100. While this captures trends among the most popular songs, it does not account for less popular songs that may explore different themes and increase the overall diversity of a genre. Consequently, low measured diversity may reflect popularity constraints rather than a true lack of thematic variation, and diversity trends might be masked because they are not observable within the top-ranked subset.

4.2 Artist Trajectory Analysis

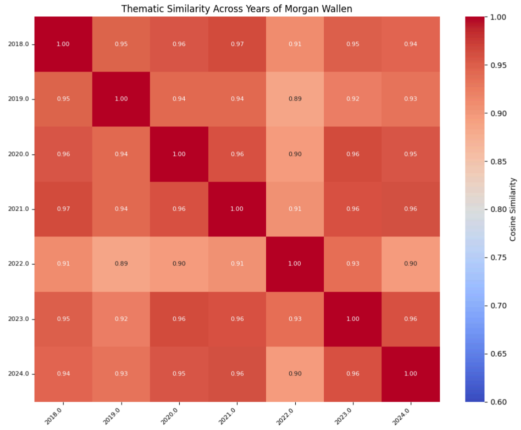
In the artist-focused analysis, attention was restricted to the three most successful artists, defined by the frequency of their appearances in the weekly Top 100 charts from 2019 to 2024. This analysis aimed at assessing whether these artists maintained a coherent thematic style across their careers or whether they introduced notable shifts in thematic direction over time. To address this, cosine similarity was computed between songs aggregated by release year, allowing for a systematic comparison of thematic consistency across years. Further, to visualize the thematic trajectories across time, the respective songs were shown in the 2D embedding space after dimensionality reduction. Together, these approaches provide insights into the extent of thematic stability and evolution in the work of each artist.

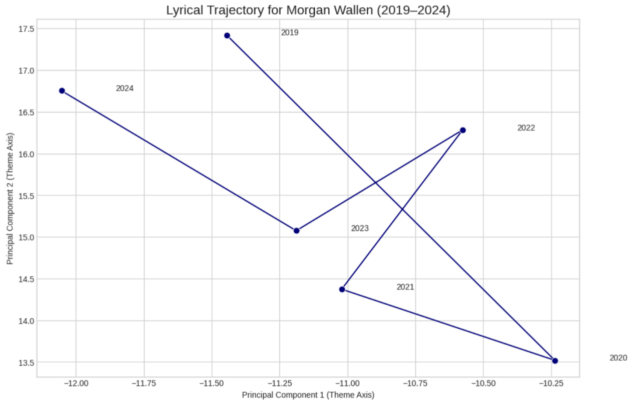
The top artist, Taylor Swift, had a lyrical trajectory showing thematic evolution and a pivot in her songwriting approach. The heatmap shows a high similarity score between the years 2020 and 2021, corresponding to her albums Folklore and Evermore. Similarly, in the PCA plot, these two years are clustered near the highest point on the y-axis (PC2), suggesting a narrative-driven and storytelling songwriting period. Then, a theme shift happens between 2021 and 2022. The heatmap shows a decrease in similarity, while the PCA plot visualizes a significant downward shift on the Y-axis. This continues through 2024, with the releases of Midnights and The Tortured Poets Department, which consist of more personal, emotional, and direct songs. However, a comparison of the charts also reveals a limitation of dimensionality reduction. While the PCA plot suggests that 2022 and 2023 are thematically close, the cosine similarity heatmap suggests a more noticeable difference. This discrepancy shows us that the 2D PCA plot is a simplification of high-dimensional data, whereas the heatmap can reveal information not seen in a low-dimensional space.



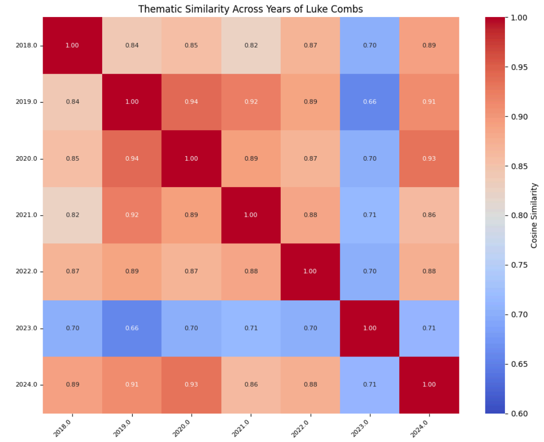


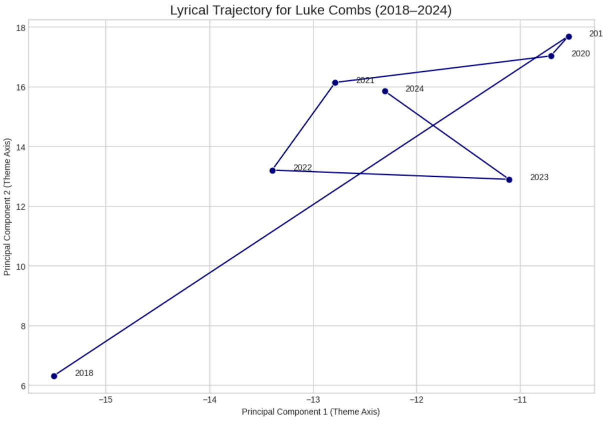
The next artist, Morgan Wallen, while more consistent than Taylor Swift, also showed volatility across the years. The heatmap emphasizes thematic similarity, with the cosine similarity between any two years of his work usually above 0.9, meaning that he has a stable, well-defined, and consistent lyrical brand. On the other hand, the PCA plot shows a zig-zag pattern, with more movement across the y-axis (PC2). This suggests that he alternates between different areas within country music. In 2019, his lyrical average is high on the narrative axis. In 2020, it goes to a more emotional and personal space, before climbing up the axis again.





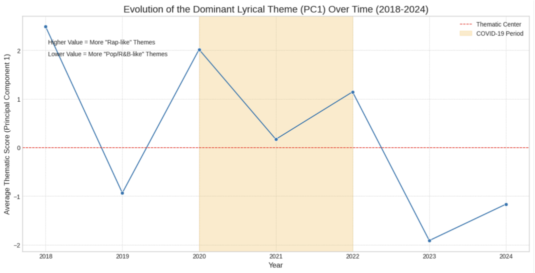
Lastly, Luke Combs also showed a clear evolution within the country sphere. The PCA plot shows a consistently high score on the y-axis (PC2), suggesting a focus on narrative and storytelling in his songwriting. It also shows a far left position on the x-axis (PC1), which emphasizes his country genre. And, the evolution from his starting point in 2018 represents his journey from traditional country music towards a more mainstream style. The heatmap shows that while most years exhibit high thematic similarity, 2023 represents a clear thematic departure, likely reflecting his album *Gettin' Old*'s more introspective and mature focus on family, aging, and personal reflection, in contrast to the upbeat, party-oriented narratives characteristic of his earlier work. This once again demonstrates how the two methods can offer different insights due to the dimensionality reduction.

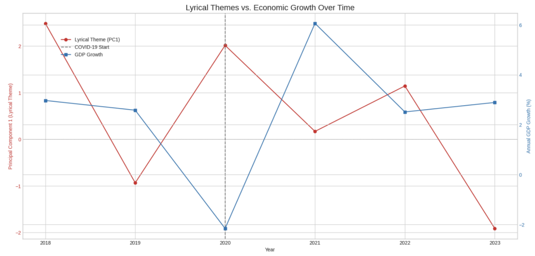


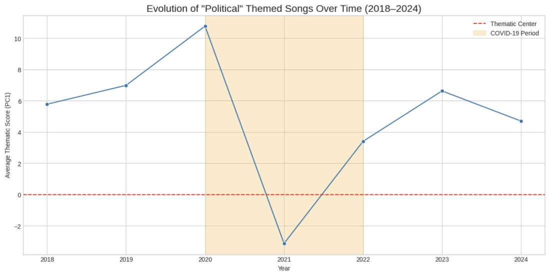


4.3 Linking Themes to Cultural Context

This section links our analysis of lyrical themes to the broader cultural and economic context. By focusing on the COVID-19 pandemic, we investigated whether lyrical themes follow economic patterns or whether they follow their own course. To begin with, in the period around COVID, the lyrical landscape of popular music swings between "Rap-like" themes and "Pop/R&B-like" themes. While the pandemic did not cause an immediate lyrical shift, there was a shift in 2022 with a move towards personal and emotional "Pop/R&B-like" themes. Next, comparing the evolution of the theme to GDP growth in the US, during the economic crisis triggered by the pandemic, lyrical themes followed their own distinct trajectory. This suggests that music's role in culture is more nuanced than simply mirroring the US economy. For instance, in 2020, when there was a sharp decline in GDP growth, the lyrical theme swung into "Rap-like" themes and later when the economy rebounded in 2021, the theme rebounded into the "Pop/R&B-like" space. This emphasizes that music is not just a reflection of the economy but rather a more complex response to it. Finally, looking at songs that mention politics or a related term, there was a small dip during the COVID period. This suggests that in a period of economic crisis and isolation, music shifted away from political commentary and more macro-related topics, and instead to more personal themes.







5. Discussion and Conclusion

The aim of this project was to explore the questions: *"How do lyrical themes evolve over time?"* and *"How do they relate to social and economic contexts?"*. During the exploratory analysis of popular music lyrics from 2018 to 2024, we found that Pop, Country, and Rap dominate the charts. Notably, Country has gained prominence over time, reflecting its growing mainstream appeal, while Rap and R&B have slightly declined in chart presence. Pop remains the most commercially successful genre overall, suggesting its broad resonance with listeners across multiple demographics. Interestingly, we observed an overall increase in positivity in songs during 2020. While this may seem counterintuitive given the predominantly negative social context of the COVID-19 pandemic, it could indicate that audiences and artists turned toward more uplifting music as a form of emotional compensation. Word clouds and LDA topic modeling revealed that *Love & Devotion* is the most prevalent theme across the dataset, with *Romantic Struggles* and *Rap/ Hip-hop* (*Expressive-aggressive Language*) also prominent. Genre-specific patterns are evident: Rap consistently emphasizes aggressive, street-oriented themes, while R&B, Pop, and Country focus more on love, personal narratives, and introspective storytelling. Despite these observations, exploratory analysis did not reveal clear overarching trends in lyrical topics over time; while some variation exists within individual genres, no consistent directional changes were found. Embedding-based analysis further clarified the thematic structure and evolution of lyrics. *Rap* topics (characterized by expressive-aggressive language) form a dense, distinct cluster, highlighting their unique lyrical style; a pattern already suggested by the exploratory analysis. In contrast, *Love & Devotion* and related topics (e.g., *Romantic Struggles*, *Heartbreak & Drinking*) are more broadly dispersed, reflecting greater thematic variety. In PCA space, the Rap-genre forms a tight, well-defined cluster, whereas other genres (Pop, R&B, Country, Rock, and Electronic) create a looser but still noticeable cluster structure, indicating shared thematic content across these genres. Overall, this confirms that for Rap, genre and topic strongly coincide, while in other genres, themes are more widely spread and multiple genres often share the same topics. Further, cosine-similarity-based distances allowed us to quantify thematic stability over time. Country and Rap show the highest stability, followed by Pop, whereas R&B, Rock, and Electronic exhibit greater variation. Interestingly, as Country became more popular, it also became slightly more diverse, suggesting that increased mainstream adoption may encourage thematic experimentation. Conversely, Rap and R&B became less popular without losing thematic diversity. Measures of diversity showed that Country and Rap are generally the least diverse, R&B and Pop the most diverse, while Rock and Electronic display the most fluctuations. Despite these variations, no clear overall trends emerged in how lyrical themes evolved during the studied period. During artists trajectory analysis we found that observable thematic evolutions often reflect personal developments in artists' lives, such as parenthood or artistic growth, rather than broad social or economic trends. Here PCA and cosine similarity analyses highlight that while dimensionality reduction helps reveal overarching trends and temporal shifts, it can also obscure subtle nuances. When considering socio-economic context, lyrical themes did not seem to strictly mirror macroeconomic trends. For example, the 2020 GDP decline coincided with a shift toward Rap-like themes, while the 2021 economic rebound aligned with a return to Pop/R&B-like themes. This suggests that music responds to cultural and societal dynamics in more complex ways rather than directly reflecting economic conditions. Interestingly, mentions of political content slightly declined during the pandemic, indicating a shift toward personal and introspective themes during periods of crisis and social isolation.

To summarize, we did not observe clear overarching trends in the evolution of lyrical themes over the studied period. Studying the impact of the COVID-19 pandemic, while there was no direct relationship between economic impacts such as GDP changes and lyrical content, we did find indications that songs tended to become more personal, introspective, and occasionally more positive, possibly reflecting societal coping mechanisms. Additionally, changes in artists' personal lives were more directly reflected in their music, and these individual shifts could be effectively captured using our embedding structure.

This work shows considerable strength and contributions. We followed an exploratory approach, validated our embedding space and explored its capabilities. Our study provides a detailed examination of lyrical themes and genres over time within a socio-cultural context. By constructing an embedding space, we were able to capture and display the nuanced relationships between lyrical content and genre. Further, we used said embedding space to conduct an extensive analysis of lyrical theme structures over multiple years while also investigating the lyrical styles and evolutions of individual artists, and investigate connections to socio-cultural and economic factors. In doing so, we extend existing research that has primarily used lyric embeddings for tagging and classification (McVicar et al., 2021; Pizarro et al., 2024), instead treating the embedding space itself as a variable worthy of investigation. However, while our work has several considerable strengths it is not without its limitations. Our evaluation setup, especially our validation set, was small and lacked diversity, which limited the interpretability of the validation loss and caused the mismatch between validation loss and improving cosine similarity accuracy. Besides that, our dataset was imbalanced across genres. While this did not prevent learning, it made training unstable. Training attempts with balanced data improved stability but resulted in worse performance overall, due to a smaller balanced set providing less training signal. Possibly using masking techniques on the balanced data could've led to better results. For validation and testing we used PCA assuming linearity which might have led to additional contextual information during dimensionality reduction. Furthermore, our study is limited by its scope. We focused exclusively on the Top 100 most popular songs in a single country and restricted the analysis to English lyrics. While these decisions were reasonable and necessary, they inevitably reduce the generalizability of our findings. For instance, additional themes may have emerged in less commercially successful songs or in other linguistic and cultural contexts, but these would not be captured in our dataset. Similarly, the absence of identifiable shifts in lyrical themes may partly reflect the relatively short time frame of the study, which may be insufficient to observe deeper cultural or economic influences on lyrical expression. Finally, our analysis concentrated solely on song lyrics, disregarding audio characteristics such as melody, rhythm, and production style, which are also central to a song's identity and cultural resonance.

Future work could address these limitations in several ways. Expanding the dataset could help with generalizability and class imbalance while also expanding the time frame of the analysis. Including audio signals such as beats per minute, energy or valence could make more nuanced analysis of songs possible. Going beyond methodological improvements, future work could move beyond exploratory analysis and investigate specific hypotheses. Fine tuning for song lyrics doesn't mean we can only embed those in the embedding space. Once fine-tuned researchers could encode textual data outside the domain of music and analyze their position and relationship to songs, themes and genres within the embedding space. This offers a promising way to investigate social, cultural and economic phenomena in relation to lyrical themes.

References

De Marzo, G. (2025, June 4). *Large language models* [PowerPoint slides]. https://giordano-demarzo.github.io/

Genius. (n.d.). *Genius API documentation*. Retrieved [August 12, 2025], from https://docs.genius.com/

McVicar, M., Di Giorgi, B., Dundar, B., & Mauch, M. (2021). Lyric document embeddings for music tagging. In *Proceedings of the 15th International Symposium on CMMR* (Online, November 15–19, 2021). https://cmmr2021.github.io/proceedings/pdffiles/cmmr2021_06.pdf

Pizarro, S., Zimmermann, M., Offermann, M. S., & Reither, F. (2024). *Exploring genre and success classification through song lyrics using DistilBERT: A fun NLP venture. arXiv*. https://arxiv.org/abs/2407.21068

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084