Evolution and Geography of Hip Hop

Parii Dan

May 28, 2024

Abstract

Lyricism has been a cornerstone of Rap since its inception, much more so than other genres. It acts as a powerful tool, with rappers expressing their emotions and opinions, discussing their failures and triumphs through wordplay and vivid imagery. Over the years, the genre has changed significantly, becoming more and more popular, there has been much debate on whether it has devolved or evolved.On one side of the debate, people claiming that presently the themes focus much more on activities harmful to the youth, glorifying things like materialism, drugs and violence, while others claim that this is just nostalgic bias, and the themes around rap have not changed as much and , at its core, they remain the same. Furthermore, Hip Hop in the United States exhibits notable regional diversity, with distinct cultural characteristics emerging from the West, East, South, and Midwest. This paper investigates the thematic and regional variations in Rap lyrics. A substantial corpus of rap lyrics is analyzed through advanced natural language processing techniques. Coreference resolution is performed using a compact model derived from the LingMess architecture. Following this, Part-of-Speech (POS) tagging and Named Entity Recognition (NER) are conducted using fine-tuned BERT models. NER tags are normalized using a novel algorithm that integrates lightweight methods. The study employs two primary methods for analyzing and visualizing the lyrics by region and decade. First, normalized NER word embeddings are visualized using Principal Component Analysis (PCA) to identify clusters. Second, the BERTopic model is applied to extract and track topics over time and across regions. This analysis aims to shed light on the evolution and regional diversity of Rap lyrics.

1 Introduction

Since its emergence in the 20th century, rap has been distinguished from the rest by its emphasis on lyricism. Unlike other genres, where instrumentals and melodies dominate, rap has been typically seen as placing spoken word at the forefront, often being compared to poetry. Through this advent, rappers can express and convey complex narratives, social commentaries, personal struggles, and accomplishments through wordplay and evocative imagery.

Throughout its existence, rap has constantly been changing. As it became more popular, moving from the boroughs of New York City to become one of the top influential genres in the world, its themes and styles have, of course, diversified a lot. These changes have sparked a debate among fans, artists and critics. Some say that modern rap glamorizes things like materialism, drugs, and violence, exerting a detrimental effect on today's youth, which is the largest demographic of rap consumption. These critics suggest a deviation and degeneration of rap, saying that the themes of struggle, social justice, and resistance have been substituted by the ones above. On the other side of the debate, others claim that these criticisms are rooted in nostalgia and the bias that comes with it. They claim that while the genre has changed and diversified superficially, its core - a medium of expression and inspiration for marginalized communities- has remained unchanged.

In addition to temporal diversity, regional variations within the US further enrich the genre and its thematic variation. Historically, there have been 4 major areas of rap [1]: East, West, MidWest, and South (Figure 1). Each region has developed its own cultural, stylistic and thematic features, influenced by its people and its economic and historical contexts.

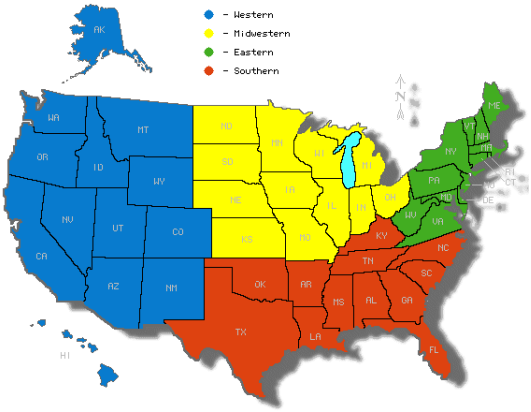


Figure 1: Regions of rap[2]

To shed more light on these topics, this study will address two fundamental questions. Firstly it will examine the evolution of lyrical themes across decades. Secondly, the research will analyze lyrics based on regional differences, investigating how distinct cultural, social, and economic contexts have reshaped the themes of rap. For clarity, we can enumerate the goals as research questions:

1. How have lyrical themes in Rap music evolved across decades?
2. What are the regional differences in lyrical content and stylistic approaches within the United States Rap scene?

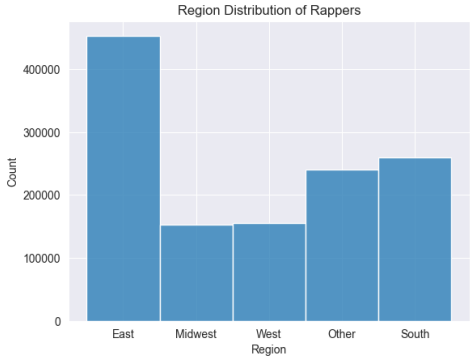
This paper endeavors to explore the temporal and regional thematic diversity within Rap music by analyzing a substantial corpus of lyrics. Initially sourced from Kaggle[3], the dataset lacked direct information on song release times and artist regions. To bridge this gap, we leveraged two API services: Geopandas API[4] to infer the rapper's region from their city and the Genius API[5] to retrieve song release times. Following data collection, cleaning, and preprocessing, we commenced the analysis phase. To achieve our objectives, we conducted POS and NER tagging using BERT[6], fine-tuned on social media data to accommodate the informal nature of the lyrics. Moreover, coreference resolution and NER normalization approaches were implemented to enhance the text further. Coreference resolution was performed using the fastcoref library [7], while NER normalization involved entity extraction, clustering through HDBSCAN [8], and pairwise comparisons utilizing the Jaro-Winkler similarity score [9].

For analysis and visualization, we employ two primary methods focusing on both regional and temporal clusters of lyrics. The first method involves converting word embeddings of normalized entities and utilizing Principal Component Analysis (PCA) to visually identify clusters. The second method utilizes the semantic power of transformer models to perform topic modeling. This method is chosen due to the nature of rap lyrics, having a lot of linguistic ambiguity, sarcasm,wordplay, references, modeling will most likely require a transformer model to do any meaningful text mining.

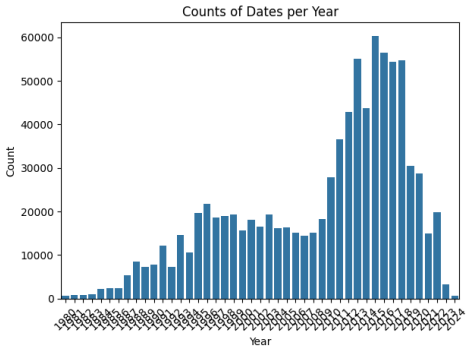
2 Data

2.1 Collection

To accomplish the goal of thematic analysis, we need a large corpus of data. Initially, the plan was to use the LyricGenius API to obtain an equal amount of lyrics from each decade, and from an equal amount of artists from our regions. However, troubles with their API service eliminate that possibility. From preliminary testing, we have established that it takes 2 minutes to extract lyrics for 10 songs. Additionally, the service tended to crash, requiring a restart to continue collecting. Moreover, the task of extracting enough lyrics per each



(a) Lyric Distribution - Region



(b) Lyric Distribution - Time

Figure 2: Lyric Distribution per class

group adds to how time-consuming this process would get.For this reason, we have opted to used Kaggle dataset[3] that already aggregated lyrics for ˜17 thousand songs. Another reason why this dataset was chosen, is because it contains the city the musician is from, which gives us a direct link to their region after further collecting.

While this dataset provides a solid foundation, it lacks the annotations necessary for our proposed analysis. These annotations would ideally link each lyric with either its region or release date. To address this, we implemented methods to extract this information. To obtain the release date, we supplied the song name to the Genius API, which returned the exact posting date. Determining the region of origin for each lyric was more time-consuming. It involved using the Geopandas API to extract geographical information on each city mentioned in the dataset, including the state it belongs to. We then classified each rapper according to their region, following the division illustrated in Figure 1. There are musicians which are not from the US,for them we have logged their country, and set them in region 'Other'.

2.2 Exploration and cleaning

Before proceeding with our analysis task, we will take a look at our data, evaluate it, and clean it accordingly. We will investigate the distribution of our lyrics concerning region and time.

The regions per lyric distribution we have obtained is somewhat skewed2, there are substantially more lyrics linked to the East Coast, with about 400 thousand out of approximately 1.2 million, but each region contains at least 150k records, so we can simply down-sample the larger ones. To get a nicer visualization, we can also plot the frequency of rappers originating from each state3. From this image, we can see high concentrations in New York, California, and somewhat Georgia. These are the 3 US states most famous for their rap scene, so it is appropriate and expected to have this kind of skew.

Examining the distribution of song release dates reveals a skew as well, albeit expectedly so. Recency bias and the growing popularity of digital platforms contribute to this skewness. For datasets not explicitly designed for time-series analyses, such skewness with respect to time is commonplace (Figure 2).

3 Methods

3.1 Processing the data

Following the collection of our complete dataset, the data still needs some processing to be used reliably in text mining for topics, or tasks similar to that. Before performing analysis,

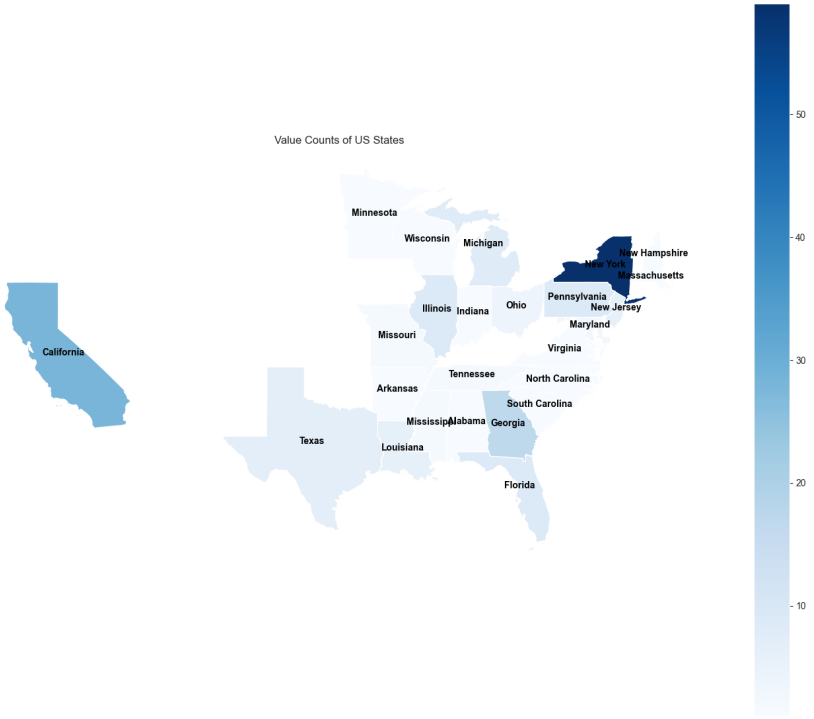


Figure 3: Plot showing the density of rappers in our data per each US state.

we have performed: POS tagging, NER tagging, coreference resolution, and normalization of entities.

POS Tagging This type of textual data has unique challenges, with rap lyrics containing intricate wordplay, cultural references, and slang. To address these challenges, we have chosen to perform POS tagging using a BERT model, which is SOTA in this domain and readily accessible. After evaluating various models and datasets for POS tagging, we found the most appropriate to be a model fine-tuned on Twitter data. This choice is based on the significant similarities between rap lyrics and Twitter data. Both are informal and rich in slang, often short and context-dense, and characterized by misspellings, abbreviations, and ambiguous language. By leveraging these parallels, our fine-tuned model is better equipped to handle the unique linguistic features present in rap lyrics.

The model outputs tokens and their corresponding classification. Further processing is required if some tokens are not full words to get a clearer view of the classifications.

NER Similarly to our approach for POS tagging, we utilized a BERT model for NER tagging as well. However, instead of finding a pre-trained model, we fine-tuned the base BERT model ourselves on the WNUT 2017 dataset. This dataset focuses on identifying novel and rare named entities in noisy user-generated text, making it particularly relevant for tasks involving non-standard language forms, such as those found on social media platforms like Twitter. Given its focus on emerging and rare entities within noisy text, the WNUT 2017 dataset is the most appropriate and accessible dataset we have found for performing NER tagging on rap lyrics. This fine-tuning approach ensures our model is adept at recognizing the diverse and evolving nature of named entities in rap lyrics.

The model generates classifications for tokens, with each token assigned to one of eleven classes: [O, B-person, I-person, B-corporation, I-corporation, B-product, I-product, B-creative-work, I-creative-work, B-group, I-group]. Here, "B" denotes the beginning of a named entity, and "I" indicates that the token is inside the entity. To produce complete named entities, additional processing is conducted to merge tokens accordingly. Following

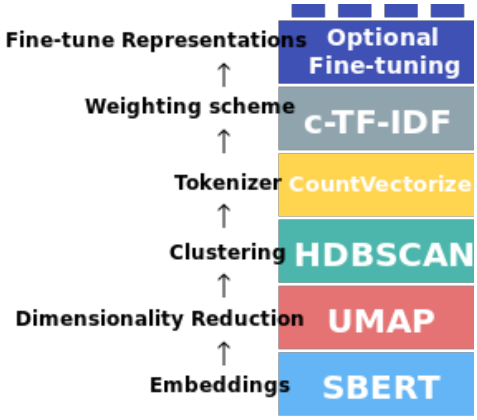


Figure 4: BERTopic Steps

this step, the output is consolidated into five classes: ['person', 'corporation', 'product', 'creative-work', 'group'].

Coreference Resolution Coreference resolution is an important processing step because it enhances text understanding by linking different expressions that refer to the same entity, thereby improving the coherence and clarity of text for further analyses. We have performed it using the compact, distilled version of a model based on the LingMess architecture [10]. This model was chosen for its speed and accessibility from the fastcoref library [7]. While the model effectively identifies coreferences, resolving them required an additional step. Initially, we planned to identify proper nouns or named entities for each coreference in a group using NER and POS models, and then resolve the reference if a suitable candidate was found. However, due to issues and time constraints detailed in section 6, we were unable to implement this approach. Instead, we adopted a simpler method: substituting each coreference in a cluster with the longest candidate from that cluster. The rationale behind this approach is that proper nouns or named entities, if present, would generally be longer than other candidates, which are most likely pronouns.

Normalization Normalization is very useful for text mining, there are many different ways to refer to the same person, product, or location. Without normalization, these might be considered different, thus harming the quality of our text. The normalization process was a longer process than the rest, as the methodology was composed of multiple different steps.

The method focuses on using the Jaro-Winkler similarity score between candidate terms. If the similarity score exceeds 0.95, the shorter word is selected as the normalized term for both candidates. This score threshold was determined through trial and error. We chose the shorter term as the substitution to handle plural nouns and different tenses effectively. However, we wanted to avoid the issue of combinatorial explosion from pairwise comparisons. With approximately 16,000 extracted "person" entities, pairwise comparison of all entities would be highly inefficient. To address this, we used the HDBSCAN clustering algorithm to group entities that are already somewhat similar. Pairwise comparisons were then conducted only within each cluster. For clarity, let's list the steps:

1. Extract Entities: Identify entities, group into lists based on their class (i.e person,product..)

2. Cluster Entities: Use the DBSCAN clustering algorithm to group similar entities together, reducing the number of pairwise comparisons needed for the next step.

3. Calculate Similarity: Within each cluster, calculate the Jaro-Winkler similarity score for each pair of entities. If the score exceeeds 0.95, proceed to next step.

4. Normalize Terms: Select the shorter word of the pair as the normalized term for both entities.

Our approach addresses two distinct cases, each with its own level of difficulty. The first case involves correcting misspellings and converting plural nouns to their singular forms. The second case focuses on identifying and normalizing a subset of abbreviations, slang, and nicknames for named entities. The reasoning behind this is that many of these substitutions tend to sound similar to the original word, such as "Drake" and "Drizzy" or "Jay-Z" and "Jigga".

Evaluation Given the fact that we do not have any annotated data, evaluating how well our model performs NER,POS tagging and normalization is not possible. The solution is to manually annotate a small subset, using more than one person to deal with bias, and then perform Kappa Analysis to evaluate agreement.

3.2 Analysis

To effectively compare lyrics based on their creation time and region, we have implemented two methods. The first method uses entities obtained from NER and normalization and the informative word embeddings obtained through BERT. The second method leverages the representation power of transformers to perform topic modeling on the processed lyrics obtained with coreference resolution.

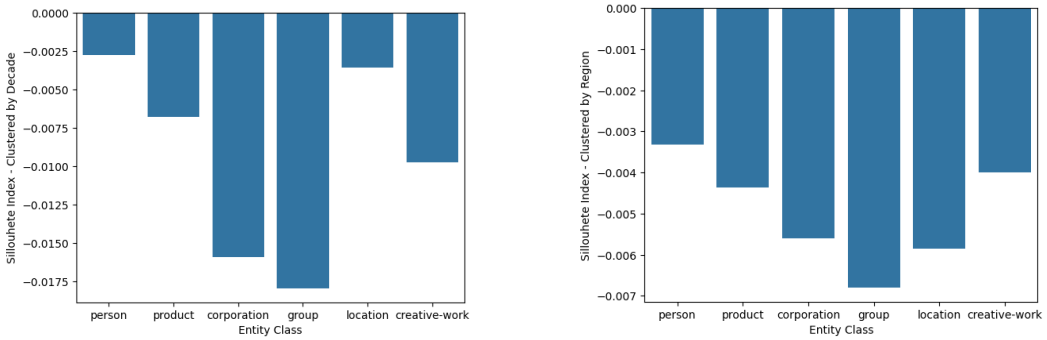
PCA on Normalized Entities The collected data is fully annotated with respect to Time and Region, enabling us to utilize these labels effectively. In this method, we process the lyrics through coreference resolution and normalize the entities. Subsequently, we generate embeddings for these lyrics using the base BERT model.We extract the word embeddings by identifying and aligning the embeddings of the tokens from the sentences. To facilitate visualization, the dimensions of these embeddings are reduced to 2D using PCA. The entities are then color-coded based on their respective Region or Decade labels, facilitating observation of regional or temporal differences in named entities.

This process is applied to each of the five processed classes identified by our NER model. Labels are assigned based on either the region or the decade of the lyric in which the entity was discovered. Besides the PCA visualizations, to further test the quality of our clusters, the silhouette index metric is applied, which measures how similar an object is to its cluster compared to other clusters.

Topic Modelling The BERTopic library facilitates topic modeling, offering a structured approach to topic representation. This algorithm consists of several modular steps(Figure 4), allowing flexibility in its implementation. Users can select different clustering algorithms, dimensionality reduction techniques, and tokenization methods to tailor the model to their specific needs.

In our customization of the base algorithm, we made several adjustments to better align with our objectives. Firstly, we employed a tokenizer model that disregards stop-words, enhancing the focus on meaningful content. Additionally, we established criteria such that a minimum of 15 words is required to form a topic, and the total number of topics is capped at 50. This limitation aids in managing the number of topics for easier exploration and interpretation.

Furthermore, we integrated seed words to guide the clustering process. Seed words are predetermined terms deemed significant to our research questions. For instance, in our case, we focused on words related to concepts such as brotherhood, equality, materialism, and drugs, which are central themes in rap music. The generation of these seed words is facilitated through GPT 4, which is appropriate for this kind of task as a generative model. By incorporating seed words, we direct the model to prioritize these themes, enriching the relevance and coherence of the generated topics.



(a) Silhouette index - Time

(b) Silhouette index - Region

Figure 5: Silhouette index per class

Method	# Samples	Kappa score	Accuracy
NER	30	0.89	0.98
POS tagging	25	0.95	0.96

Table 1: The Kappa scores and Accuracy of the human validation process, on a small subset of samples

Class Imbalance In our data collection process, we've encountered an imbalance in the distribution of records across different regions and decades, as detailed in section 2. Ideally, we would strive for an equal number of samples in each class to ensure a balanced representation. However, given time constraints, we've opted for a more realistic and biased approach to address this issue.

To deal with the imbalance, we've employed a combination of downsampling the overrepresented classes and slightly upsampling the underrepresented ones. This strategy aims to achieve parity in sample size across all instances while efficiently utilizing the available data. Although it may not perfectly rectify the imbalance, it provides a practical compromise within our constraints, facilitating more robust analyses and interpretations of the data.

4 Results

Following the procedures outlined in section 3, we now have a variety of methodologies to evaluate the results. Firstly, we will assess the performance of our strategies for Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and text normalization using Kappa Analysis. Subsequently, we will delve into the examination of metrics and the visualization of the PCA-Embedding approach. Finally, we will delve into the exploration of compelling topics generated through the Topic Modeling method. This structured approach ensures a comprehensive analysis of our results across different techniques, enabling a thorough understanding of our findings.

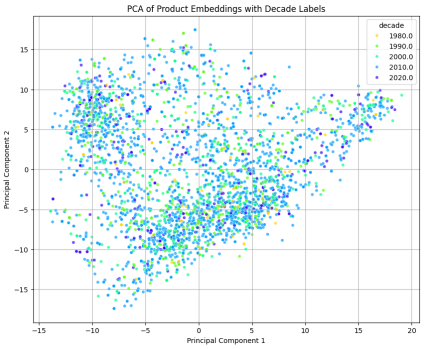
4.1 Kappa Analysis

The NER and POS model results are presented in Table 1. While scores for both kappa and accuracy are high, it's important to note that our assessment is based on a limited subset of 25-30 samples. Consequently, we cannot confidently assert the model's performance based on this small sample size.

4.2 PCA-Embedding

For this method two evaluation approaches are outlined: Silhouette index scores and PCA visualizations.

(a) PCA - NER 'person' labeled by decades.



(b) PCA - NER 'product' labeled by decades.

(c) PCA - NER 'person' labeled by regions.

Figure 6: PCA - Entity Class v Region/Decade

A set of 10 silhouette index scores were computed, each representing a different combination of entity type and label type. Figure 6 displays these scores. They uniformly exhibit low values, indicating that our predefined clusters, whether based on region or decade are not distinctive at all.

We've extended our evaluation by computing PCA for all (entity type - region) and (entity type - decade) combinations, including subsets such as the West and East coast. Below, we'll present a condensed sample of these outcomes to offer a broad perspective on our findings.

1. Person - All decades (Figure 6a): In this visualization, we present our dimensionality-reduced person embeddings, with each point color-coded according to its respective decade.The plot reveals no discernible patterns or distinctions across different time periods.

2. Product - All decades (Figure 6b): Our aim here is to explore potential differences in the products mentioned in rap lyrics across the five decades. While the plot displays two somewhat distinct clusters, they fail to align with our predefined labels.

3. Person - East v West (Figure 6c): Transitioning to regional analysis, we now compare named individuals from the East Coast and the West coast. Once again, the visualization fails to reveal any notable distinctions between the two regions.

4.3 Topic Modelling

The topic modeling approach has effectively grouped lyrics into expected themes. It accurately identified topics such as materialism, regionalism, racism, and various forms of word-play.

The results of the topic modeling approach will be divided into two sections: Topics Per Region and Topics Over Time.

Topics Over Time The Dynamic Topic Modeling approach is set to group lyrics into 50 themes. In our exploration, we'll delve into several interesting ones with respect to their change over time. On the right-hand side of the displayed plots, you'll notice the selected topics from our pool. While the model automatically generated the labels, we took care to verify their appropriateness by examining specific instances, ensuring they align, at least generally, with the identified themes. We will display a line plot of the frequency of use, along with a wordcloud plot that shows the most representative words of the.

1. Money Over Time (Figure 7a): The plot indicates a peak at the beginning of this decade, as well as a much smaller peak in the late 80s.
2. Race Over Time (Figure 7b): The text associated with the topic of racism is quite different from the money one, it is almost flipped.
3. Education Over Time (Figure 7c): The topic of education throughout time has peaks in mid-to-late 80s and a much smaller peak after 2020.
4. Crime Over Time (Figure 7d): The frequency of the topic crime is almost symmetric, there are two similarly sized peaks in ,once again, the mid 80s and early 20s.

There are some clear distinctions over time of these topics, in this specific context, with this data and the topic modeling approach. More on this will be discussed in section 5.

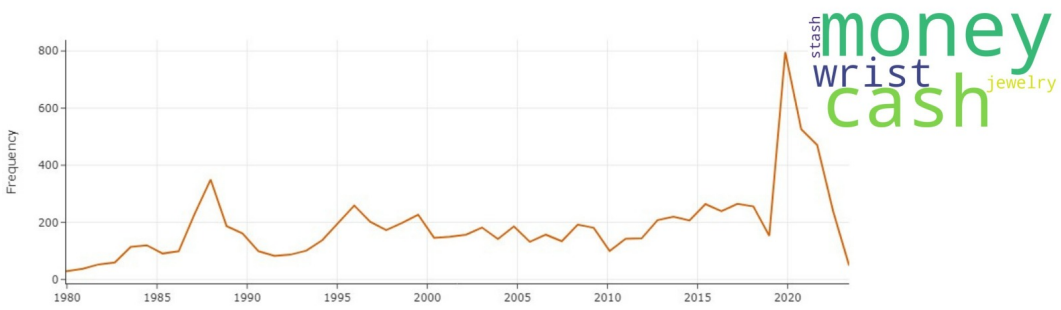
Topics Per Region The results of topic division by region are not as distinct as those over time. For the sake of brevity and clarity throughout the paper, the plots displaying these results are included in the Appendix. Figure 9 shows that the four most prominent topics—focused on money, drugs, and more abstract themes like emotion and optimism are relatively consistent across all regions. However, some regional distinctions do emerge. In Figure 10a we can see that the West seemingly uses more Spanish slang in their lyrics. However, this may be due to the model misinterpreting "LA" as the Spanish article "la" instead of referring to "Los Angeles." Figure 10b, shows us that the West Coast mentions jewelry less than other regions.

5 Discussion

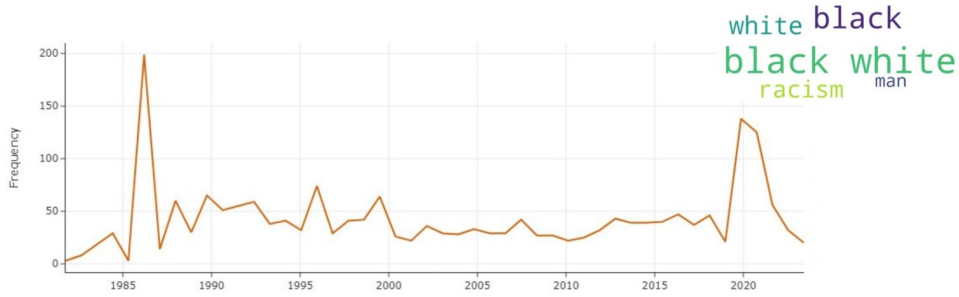
In this section, we will discuss the results obtained by each of our methods, and attempt to clarify some aspects, as well as propose future investigations and corrections to this process.

Processing Methods The NER and POS tagging results are a positive sign with respect to the respective models' capacity in these two tasks. Nonetheless, it's crucial to acknowledge the limitations imposed by our small subset of data. Drawing definitive conclusions from these results is not possible. In order to expand on this, further tests need to be performed on larger labeled datasets to validate and refine the models' performance.

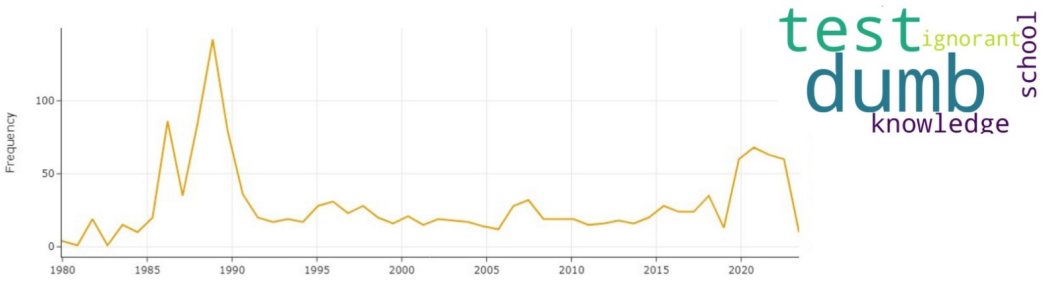
Analysis We began our evaluation with the PCA clustering approach, utilizing BERT embeddings generated from the normalized named entities in our processed data. These embeddings were then labeled according to either region or decade. Subsequently, we assessed the quality of the labeled clusters using both the Silhouette Index and 2D PCA visualizations. Unfortunately, the results were underwhelming, failing to reveal any discernible distinctions. There can be multiple reasons for this: 1)The entire set of normalized entities was not of enough quality to facilitate this kind of clustering in further steps, 2) The embeddings generated by BERT did not capture enough meaning to have entities from similar decade/region be similar. One other reason can be the fact that entities are naturally



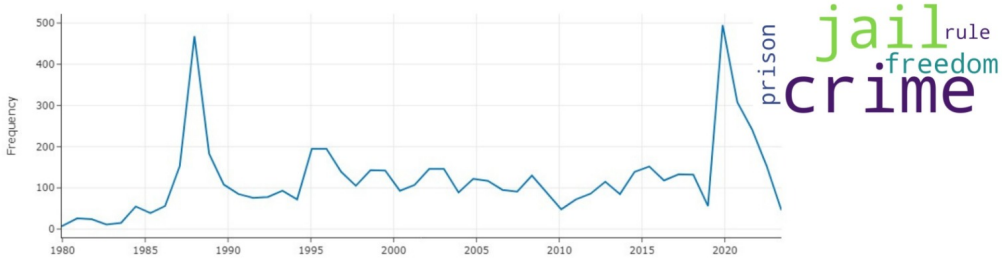
(a) Topic Over Time - Money



(b) Topic Over Time - Race



(c) Topic Over Time - Education



(d) Topic Over Time - Crime

Figure 7: Dynamic Topic Modelling

grouped by aspects besides their time of origin. For example, in the product plot 6b, it is very possible that one of the big clusters is something like cars, and another shows jewelry or another type of product. So, to effectively perform this analysis, an idea can be to further separate something like the 'product' named entities in more specific classes, and then apply the method to those.

Transitioning to the Topic Modeling methods, we found more pronounced distinctions

and interesting findings. However, before delving into a discussion, it's essential to address several important considerations. Firstly, the dataset we collected represents only a small subset of all rap songs released over the five decades. Moreover, the distribution of data across regions and, particularly, over time is skewed, with a notable abundance of recent data compared to the 80s and 90s.To address this skewness, after exhausting other collection possibilities, we adopted a simple approach by sampling with replacement to ensure equal representation across classes. However, it's crucial to recognize that this method may introduce unaccounted bias, potentially skewing our results. The frequency plots of the temporal topic data indeed reveal underlying distribution patterns, with two peaks observed on each side on almost every plot (Figure 7). Despite this, further examination of the sample distribution per decade confirmed equal representation, as initially intended (Figure 8).As such, at present, we cannot definitively determine whether our data is genuinely biased or if these topic peaks naturally occur. Further investigation and analysis are required to provide more clarity on this matter. However, even if the data is not unbiased, the distinctions are clear enough to indicate some definite changes in topics throughout the region and, especially, time.

We observe a notable prevalence of positive themes such as education in the past compared to the present, with a significantly higher frequency. Upon examining the specific lyrics classified under this topic, we find instances where education is depicted neutrally rather than positively, as seen in phrases like "and they caused mad problems like math exams." However, there are numerous examples portraying education in a positive light, such as "you should have gone to school you could've learned a trade," as well as negative portrayals like "and the curriculum be tricking them dollars I spend." Consequently, further investigations are necessary on a per-topic basis to accurately quantify these changes. In the case of education, however, based on the samples analyzed, the majority conveyed a positive or neutral message about school.

Conversely, negative and materialistic themes like money are more prevalent in the present compared to the past (Figure 7a). However, themes associated with worse societal issues, such as "crime," "jail," and "prison," display a peak in the late 80s, followed by a gradual decline and another peak in the early 2020s(Figure 7d). Establishing a definitive correlation in such complex scenarios is challenging.

Words related to race, white and black people specifically, are also noticeably more prevalent in the past, with a gradual decline followed by another peak in recent times. Once again, however, we cannot definitively tell whether the sentiments expressed are mostly positive or negative, we can simply mention their prevalence.

In addition to the temporal analysis, we conducted region-based analyses. However, this approach yielded less insightful findings, as there were no significant changes observed across the different regions. The four major topics of discussion, including themes of religion, freedom,drugs, and crime were uniformly distributed throughout the regions.

Research Goal Our initial goal for all of this analysis was to answer two specific research questions. The first one seeks to establish how lyrical themes have changed throughout the years, while the second one is focused on themes and approaches distinguished by region in the US.

1. Our analysis revealed that positive themes such as education and equality were more prominent in the past, though they still persist to some extent in contemporary music, albeit to a lesser degree. Conversely, negative themes such as materialism and money are notably more prevalent in today's music landscape. Lyrics focusing on crime and prison are prevalent across the years, with peaks in the late 1980s and early 2020s. However, it's important to note the limitations of our approach, particularly regarding the skewed distribution of our time data. Introducing sampling with replacement may have introduced bias, affecting the reliability of our findings.

2. The comparison of lyrics from different geographical origins did not yield distinctions as pronounced as those observed in the temporal analysis. Across the top five themes,

which represent a significant portion of our dataset, similarities in frequency were evident among regions. It appears that, particularly for prevalent themes like "money," "crime," and "religion," rappers from diverse regions discuss them in similar proportions.

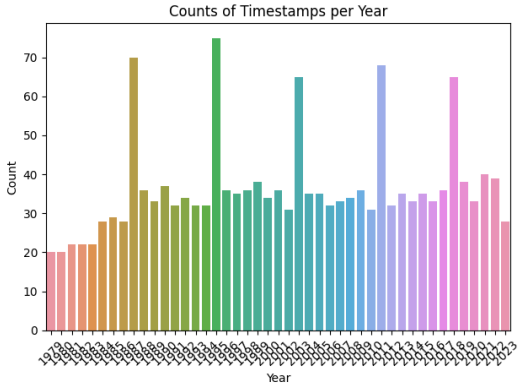


Figure 8: Distribution of time following Sampling with replacement. We can observe here that there is not dominance in either the mid-to-late 1980s or the early 2020s.

6 Challenges and Considerations in Research Implementation

Throughout this project, several challenges arose, significantly impacting the planned workflow. One primary hurdle stemmed from the time-consuming nature of certain processes, particularly those involving BERT, given the large corpus of 1.2M rows of lyrics. Moreover, reliance on online platforms like Kaggle and Colab introduced additional complications, such as running out of credits or encountering disconnections due to inactivity or CUDA dependency issues. These disruptions necessitated frequent restarts, consuming valuable time and hindering progress. Consequently, some aspects of the project, like Sentiment Analysis and Negation Handling on lyrics grouped by topic, were omitted due to time constraints.

Another significant issue was the inability to obtain an equal amount of data across different time periods or regions. The planned use of the LyricGenius API for data collection proved inefficient, with frequent timeouts leading to prolonged data collection times. As a result, we resorted to utilizing a Kaggle dataset, which lacked the class balance originally intended for the analysis. This disparity introduced uncertainty into the results, impacting the robustness and reliability of our findings.

7 Ethical Considerations

The discussion around the ethics of this study is crucial. Racism persists, and although its prevalence has diminished, many people still hold racist views. These individuals might draw false conclusions if they superficially analyze the results of this project, assuming that modern rappers are primarily focused on money and crime. This is a gross misrepresentation of the complexity and genuine artistry present in rap music. It's essential to reiterate that our dataset represents only a small subset of all rap songs, and the genre as a whole. Furthermore, our data was skewed and sampled to ensure class balance, which is not an ideal solution and can perpetuate existing themes and biases.

8 Concluding remarks

The project has uncovered intriguing insights into the temporal and regional variations within rap music. However, further examination is warranted to validate and replicate these findings, especially using less skewed data to ensure robustness. Additionally, an avenue for future research involves incorporating Sentiment Analysis into our Topic Modeling approach. This addition would enable us to quantify the positivity or negativity of the text grouped within each topic, providing a more nuanced understanding of thematic changes over time.

References

[1] Bradley Piri. *The Geography of Hip-Hop — B. Piri Photography*. Feb. 2021. URL: https://www.bdotpiri.com/blog/hiphopgeography.

[2] *Best Hip Hop region*. URL: https://genius.com/discussions/343549-Best-hip-hop-region.

[3] Jamie Welsh. *Rap Lyrics*. Sept. 2023. URL: https://www.kaggle.com/datasets/jamiewelsh2/rap-lyrics.

[4] Kelsey Jordahl et al. *geopandas/geopandas: v0.8.1*. Version v0.8.1. Apr. 2020. DOI: 10.5281/zenodo.3946761. URL: https://doi.org/10.5281/zenodo.3946761.

[5] Johnwmillr. *GitHub - johnwmillr/LyricsGenius: Download song lyrics and metadata from Genius.com*. URL: https://github.com/johnwmillr/LyricsGenius?tab=readme-ov-file#readme.

[6] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[7] Shon Otmazgin, Arie Cattan, and Yoav Goldberg. "F-coref: Fast, accurate and easy to use coreference resolution". In: *arXiv preprint arXiv:2209.04280* (2022).

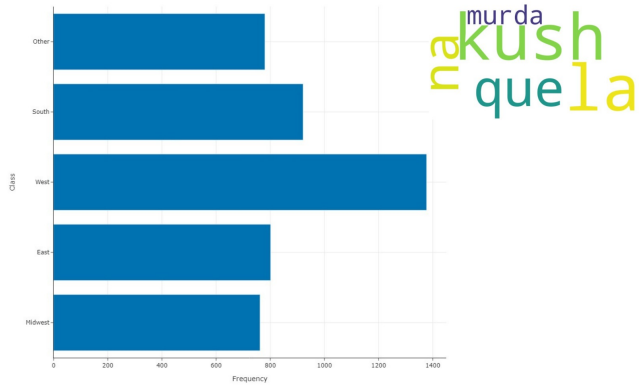
[8] Erich Schubert et al. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN". In: *ACM Transactions on Database Systems (TODS)* 42.3 (2017), pp. 1–21.

[9] Yaoshu Wang, Jianbin Qin, and Wei Wang. "Efficient approximate entity matching using jaro-winkler distance". In: *International conference on web information systems engineering*. Springer. 2017, pp. 231–239.

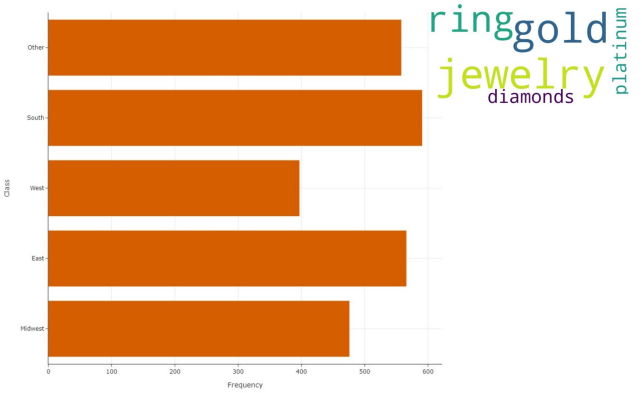
[10] Shon Otmazgin, Arie Cattan, and Yoav Goldberg. "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution. CoRR abs/2205.12644 (2022)". In: *URL: https://doi. org/10* 48550 (2022).

9 Appendix

Figure 9: Topic Modelling Per Region - Biggest 4 Topics (Emotion,Pain/Fear,Strees/Ghetto,Dream/Time,Light



(a) Topic per Region - Spanish Slang



(b) Topic per Region - Jewelry

Figure 10: Topic Modelling Per Region