Analyzing song lyrics while comparing LDA and BERTopic in terms of Topic Modelling

Maurice Droll (maudr877)

Abstract

This paper contains a song text analysis, which compares LDA and BERTopic to model the topics for song lyrics between 1970 and 2025. The dataset was chosen from the genius.com website and was preprocessed. Various hypotheses were evaluated, for example whether love is a permanent theme in song lyrics. The results of both approaches were compared and the influence of Sentence Transformers was determined. BERTopic performs significantly better than LDA and the Sentence Transformers have little influence on the selected question.

1 Introduction

"Song lyrics have become simpler and more repetitive over the last five decades" Parada-Cabaleiro and Mayerl, 2024 p.1, which offers the perfect opportunity to find out the main topics with topic modeling.

Additionally, songs sound subjectivly more and more the same and face subjectivily the same topics. This paper explores the question of whether the same themes always appear in songs and whether trends can be recognized over time.

Furthermore, the question being addressed is whether it is sufficient to use LDA or whether the analysis should be done with BERTopic. This processing could also be used, if the estimation is good enough, to estimate further music industry trends and even provide the possibility for artists to enhance their lyrics by checking with the model like this one for popularity.

2 Theory

2.1 What is Topic Modeling?

Topic Modeling describes a technique to pick up represented topics in a text or a group of texts and therefore using probabilistic calculations. A topic can be seen as a theme or an underlying idea, which was represented in document Srinivasa-Desikan, 2018 p.1.

2.1.1 What is BERT?

BERT (=Bidirectional Encoder Representations from Transformers) is using a pretrained neuronal network to form an understanding of context of words in texts. It relies on bidirectional processing and Masked Language Modeling (MLM), where random parts of the sentence are masked and the model predicts the words based on the context before and after it. Additionally it used Next Sentence Prediction to understand relationships between sentences. All is built on a Transformer architecture Devlin et al., 2018 p.3-4.

What is BERTopic? BERTopic was developed by Maarten Grootendorst and is "a topic model that extends this process by extracting coherent topic representation through the development of a class-based variation of TF-IDF" (p.1). The topics get represented in each cluster, where each cluster has one topic. It uses cluster algorithms such as HDBSCAN. Additionally, an class based TF-IDF algorithm analyses, which words are most relevant. The implementation allows to use different sentence transformers and therefore allows for flexibility for the use case and the available hardware (CPU or GPU based). In the papers intern benchmarks BERTopic showed a significant improvements in terms of speed, topic coherence and topic diversity compared to other typical topic modeling approaches like LDA (Grootendorst, 2022 p.6f.). [1]

2.1.2 What are Sentence Transformers?

Sentence Transformers are being used for encoding sentences into high-dimensional vectors. This can be utilized for semantic search, clustering and sentence similarity. They can be seen as a deep learning model to understand the semantic meaning of sentences, while exceeding the traditional

[1] https://maartengr.github.io/BERTopic/index.html

capabilities of traditional word embeddings Team, 2024.

2.1.3 What is LDA?

As seen in the lecture, LDA (= Latent Dirichlet allocation) can be used in order to extract topics from documents in a statistical way. A topic is defined by the probability of words Blei et al., 2003, p.996. LDA is a common tool for identifying topics in large amounts of text.

2.1.4 What are BiGrams?

BiGrams are sequences of word in a text. They are usually identify word pairs like artists or famous people. In this given example it could be common sayings or artists like Justin Timberlake Schonlau et al., 2017 p.867.

3 Main Part

3.1 Data

The song lyrics were being used from the Genius.com website [2], which allows users to look up song lyrics of their favorite artists and comment on the lyrics. This data set was chosen over comparable data sets on the basis of good personal experience and a large selection of song lyrics from many genres and times. The data set also contained the columns 'views' , 'artist', 'tag' = 'genre' and 'language', which other datasets did not offer and what would result in an enormous manual annotation effort of all 2.76 million songs. Also, it covers songs from the 1970s until today and does offer a views counter for the popularity of the songs. Therefore, this is suitable for the selected application.

3.1.1 Preprocessing

Preprocessing is an important step. Lots of crashes have been caused by special characters, while BERTopic or LDA analysis. Since there are a lot of songs available, manual skimming through every song text is impossible. The data set is much less standardized compared to data sets from production devices, e.g. an industrial saw with Industry 4.0 IoT monitoring.

Therefore, a lot of guidelines and preprocessing steps were established:

1. Reduction to English-language songs

   The dataset is multilingual; however, due to better comparisons and less mix-ups while analyzing, only English song texts were used. The decision was easy to implement as there was a column with the language used. A sub data set has been formed for every genre [rap, pop, misc and rock] and every decade. This was done using heapq and saving in on its own to the hard drive to reduce the processing time. For every decade only the most viewed 1000 songs of each genre and decade have been used.

2. Sorting out unnecessary characters

   For easier analysis certain non-alphabetical characters such as brackets have been sorted out. Especially in rap music, a lot of apostrophes are being used, which have no influence on the topics of the song, but can change the result of the calculation. Therefore, the regular expression 'r"[a-zA-Z0-9]"' was applied.

3. Deleting stop-, swear- and formatting words

   To make the analysis easier, common words to format have been deleted such as '[Intro]', '[Hook]'. Additionally, since a lot of rap and pop music uses swear words, they have been removed. If replaced by "swear", this have turned out as the most relevant topic. The swear words were important from two English data sets and supplemented by own findings. The data sets are available at kaggle. [3] [4] Also the English stopwords of nltk have been deleted.

4. Lemmatization

   Lemmatization is used in order to improve the analysis, by limiting word variations. The analysis does not get affected if the song lyrics contains "organize" or "organizing".

5. Avoiding collisions

   The analysis of each genre has been organized into multiple jupyter-notebooks to avoid collisions.

After preprocessing, the song lyrics were stored in an array and processed with BERTopic, LDA and BiGrams. For easier analysis the results were saved in .csv files.

[2] https://huggingface.co/datasets/sebastiandizon/genius-song-lyrics

[3] https://www.kaggle.com/datasets/tushifire/ldnoobw

[4] https://www.kaggle.com/datasets/sahib12/badwords

3.2 Method

The hypotheses that will be tested later are presented below.

3.2.1 Music deals with political topics and relevant topics to people care at the moment

Popular songs often face important political or important social issues. To investigate this, an online source was consulted to compare the most important terms with the topics identified Media, 2024.

3.2.2 Love songs are and were always important

Studies have shown that music and love are connected together and that love has been a frequent topic in music for centuries and this is regardless of the culture of the people Bamford et al., 2024 p.1. If this thesis is correct, the most frequent topics of the last 6 decades should include a large number of topics relating to the overarching theme of love.

3.2.3 Newer Songs are more repetitive than older ones and have fewer different topics in them

"Song lyrics have become simpler and more repetitive over the last five decades" Parada-Cabaleiro and Mayerl, 2024 p.1. In order to try to substantiate the thesis, the agreement values (=weights) of the topics are compared and checked. If the values increases, this indicates an increase in the focus on fewer topics.

3.2.4 Artist and other influential persons can be found in terms of BiGrams

The press occasionally reports that artists name each other in their songs Staff, 2024. Taylor Swift and Justin Bieber and other famous artists should be part of music lyrics. The press occasionally reports that artists name each other in their songs. This means that well-known artists who, in addition to their music, also provide social discussion material (e.g. Tailor Swift or Justin Bieber) should also appear in song lyrics. This will be investigated with a BiGrams search and BERTopic.

3.2.5 Famous events are also part of songs lyrics

The same should be true for reoccurring periods in the year such as Easter and Christmas. BiGrams are also analyzed for this purpose.

3.3 Results

3.3.1 Music deals with current political and otherwise relevant topics

In both sankey charts (LDA/BERTopic) there is no clear indication of frequent, clearly political statements. The most frequent topic, which appeared in the files is war and love. Also since a lot of political song writer were mentioned it can be estimated that political statements are part of many song (examples: Tupac Shakur, Eminem).

3.3.2 Comparison between different Sentence Transformers

Sentence Transformers differ in their architecture and semantic accuracy. They are therefore differently suited to the deployment scenario. Since there is no obvious choice of a sentence transformer for song lyrics analysis, the most common transformers with a slightly different focus were chosen. Two models (MiniLM and all-Mpnet-base-v2) with the same focus were also selected. Particular emphasis was placed on execution time and any unforeseen differences.

There are mainly technical differences between the various SentenceTransformers, but few differences in terms of content. As shown in the table, some of the themes are somewhat narrower. They also differ in number. For example, the topics with LaBSE are very well summarized, which means that important details may be omitted and certain statements, such as about the relevant artists, may not be possible because they do not appear in the topic illustration. LaBSE and Paraphrase-Mpnet-base-v2 are suitable if you are looking for specific topics or have made a strong restriction in this regard.

In the selected setting, the restriction would be waived. This leaves the Sentence Transformer MiniLM and All-Mpnet-base-v2. They are balanced in terms of detailed analysis and diversity. MiniLM was chosen for further analysis, since the processing time is shorter. MiniLM took 534 seconds and All-Mpnet-base-v2 needed 2230 seconds for analyzing 1000 songs per decade. Compared to LDA both sentence transformers work faster, since LDA needed more than 4 hours, which is equal to more than 14400 seconds. The comparison is also listed in table 1.

3.3.3 Newer Songs are more repetitive than older ones and have fewer different topics in them

The following section analyses, whether the most common themes in songs become more trivial over time. In terms of Pop, the Figure 6 shows a word cloud from 1970s - 2020s. The bigger the word, the important the "weight" of the token. Below it is a scale that indicates the weight of the topic in colours. In the 1970s it focuses on Merry Christmas and a focus on love, in the 1980s focus on a verse a love and baby as fashion word. In the 1990s it shift to love and religion (jesus). One decade later it mainly emphasizes two important artists: Troy and Justin Timberlake. Those important artist get replaced by Harry Styles and Liam Payne one decade later. The 2020s have no clear topics.

The result in RAP are similar (Figure 7). The rap genre was still developing and later on if focused a lot on important artists (Tupac Shakur, Eminem, Taylor The Creator) and typical rap words like "yeah".

The result in Rock are also similar to RAP (Figure 8). It features famous persons like Roger Waters and contains the era of Chester Benningfield and Linkin Park as well as Machine Gun Kelly.

There is a recognizable trend for topics to be concentrated on just a few. Individual artists are highly valued. Sometimes, however, there are also effects, such as with the Machine Gun Kelly , whereby "gun" is also found as an independent word.

3.3.4 What is the most important topic over all genres and decades?

As shown in Figure 4 the most important topic is "im" (2.16), followed by "christmas"(1.02) and "love" (0.827) using BERTopic. "Im" was performed in all decades except 1970, which speaks in favour of a strong ego reference in musical works and emphasizes the importance of one's own person. "Christmas" is influenced mainly by 1970s and 1990s and is not as relevant in other decades. There are other conspicuous features, for example artists such as Chester Bennington (2000s rock) or Gil Scott-Heron (1970 rap) are listed as separate topics and thus as presumably major influencing factors.

Per decade and genre the most relevant topic were added and therefore the most relevant topic was created.

The second analysis by LDA shows that love (0.56) is the biggest topic, follow by rock (0.154) and time (0.112). So one can see that the topics are more straight forward and from every decade there is a big influence on the topic of "love". However the overall added value of love is only 0.56, which shows the weight of the topic is much lower than using BERT.

3.3.5 Love songs are always important

So as shown in the LDA sankey chart (5, "love" is relevant in all decades. When looking at the "pop" and "rock" genre, "love" is the number one topic in every decade. The most genre that changed the most was "miscellaneous". It showed the biggest variety ranging from "Monica" and "Ross" (1990s) to "people" and "time" (2010s). Monica and Ross are most likely fictional characters of the Friends TV Series (1994- 2004), which would fit in the decade. Rap music is mostly about "dreams" (1990s), "play" "game" (2000s) and "rich" and "money" (2010s).

3.3.6 Using famous artists in the own music gets more popular over time

Unfortunately, no BiGrams referring to artists could be found in the song lyrics examined. However, BERTopic was able to pick up a lot of artists in Pop, Rap and Rock. This can be visualized in the bar char for rap (Figure 2) , pop (Figure 1) and rock (Figure 3).

For those decades are those artists analysed as topics: 70s (Eva Perón, Che Guevara), 90s (Justin Timberlake), 10s (Harry Styles and Liam Payne) and 2020s Taylor Swift to name a few examples. In terms of rap there are lot of artists and famous person involved, like illustrated in the figures 2 3.

Finally, the result is that the thesis cannot be accepted. The use of artists in song lyrics is at its highest level between 1990 and 2010. If the thesis would be true, the usage should continue to increase and not get lower again.

3.4 Discussion

3.4.1 Correct and incorrect hypotheses

First of all, song lyrics mostly contain love, but love is not the most relevant topic in every genre. The usage of famous artists in the own music does not get more popular over time. Songs lyrics oftentimes rely on artists as their main topic. Utilization has neither increased nor decreased significantly. The level and the statement that music is becoming increasingly monotonous in terms of content could

not be confirmed. Also, the influence of the Sentence Transformers was minimal. Therefore the commonly used MiniLM was used here also.

3.4.2 Comparison to scientific articles

In a paper called "Exploring Genre and Success Classification through Song Lyrics using DistilBERT" Martinez et al., 2024, a special version of BERT is being used to classify song lyrics to one of five categories. Using BERT and Support Vector Machines, it was possible to predict the year of release. Compared to the approach of this paper, songs with less than 100 words were sorted out and authors implemented an own feature extraction model. The paper claims that nearly 80 percent of songs can be predicted correctly, if the song is successful or not. This shows the huge potential of BERT. The paper "Mining Sentiments from Songs Using Latent Dirichlet Allocation" Sharma and Murty, 2011 uses LDA to recognize the mood of a song. The preprecessing contains also removing of stopwords and tokenization. Topics are than being matched to emotions. The paper can be seen as possible part of a music recommendation system.

3.4.3 Limitations

First of all, hardware limitations did decrease the amount of song to be processed. An LDA processing of 4000 songs per decade took around 4 hours. Even running 100‘000 songs on faster hardware, would have caused inappropriate amount of time. BERTopic was much faster, however in order to compare it, it has to be the same amount of data. Also manually sorting out songs would improve most likely the result, however due to time constraints was waived. Especially when it comes to seeing the view counter as a means of popularity of the songs. The lyrics of the most interesting songs for the majority of the people is being looked up on platforms such as genius.com. So songs, which have the most amount of streams on platforms like Spotify or Amazon Music may better represent the most important songs of the time. Nevertheless, this decision was made deliberately, as it can also be assumed that these songs have more content.

4 Conclusion

When analyzing song lyrics, it is important that the results have a certain depth. This depth could be achieved by using BERTopic, but only in part by using LDA. In principle, relevant topics can be analyzed with LDA, e.g. ‘love’, but less relevant artists of the time or important political topics. In general, BERT proved to be a more suitable approach.

Due to the limitation to 1000 songs per decade and genre, it would be advisable in future to analyze the complete data set of sometimes over 100,000 songs per decade and genre. In particular, topics that were not picked up by the most streamed artists could then be given more influence in the analysis. So investing more time into preprocessing and even performing a manual preselection, which not only relies on streamed views, can improve the quality quite a bit.

Additionally, it might be beneficial to use some sort of summarizing functionality of a Large Language Model to decrease the amount of data processed and maybe to set the focus more on the content instead of the musical subtleties and filler words that should not be taken into account in the implemented procedure anyway.

With this review, hypotheses regarding the relevance of love in music, general themes and the use of artists and musicians could be proven. However, some of the above hypotheses could also be refuted.

The results of the analysis are not good enough to recognize recurring trends or predictions for suitable songs of a decade so that these can be used in song composition. If such projects are to be implemented, it is advisable to build on a transformer-based approach and thus profit from the semantic connections and recognized emotions. If one only wants to achieve pure topic clusters without any particular depth, then LDA can also be used as well.

References

J. S. Bamford, J. Vigl, M. Hämäläinen, and S. H. Saarikallio. 2024. Love songs and serenades: a theoretical review of music and romantic relationships. *Frontiers in Psychology*, 15:1302548.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Servando Pizarro Martinez, Moritz Zimmermann, Miguel Serkan Offermann, and Florian Reither. 2024. Exploring genre and success classification through song lyrics using distilbert: A fun nlp venture. *Preprint*, arXiv:2407.21068.

Triangle Direct Media. 2024. Top google searches and phrases by decade. Accessed: 2025-01-20.

Elena Parada-Cabaleiro and Matthias Mayerl. 2024. Song lyrics have become simpler and more repetitive over the last five decades. *Scientific Reports*, 14:5531.

Matthias Schonlau, Nick Guenther, and Ilia Sucholutsky. 2017. Text mining with n-gram variables. *The Stata Journal*, 17(4):866–881.

Govind Sharma and M. Narasimha Murty. 2011. Mining sentiments from songs using latent dirichlet allocation. In *Advances in Intelligent Data Analysis X*, pages 328–339, Berlin, Heidelberg. Springer Berlin Heidelberg.

B. Srinivasa-Desikan. 2018. *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, SpaCy and Keras. Packt Publishing Ltd., Birmingham.,*. Packt Publishing Ltd., Birmingham.

AltPress Staff. 2024. Bands referencing other bands in songs. Accessed: 2025-01-20.

Marqo Team. 2024. Introduction to sentence transformers. Accessed: 2025-01-20.

A Appendix

A.0.1 Artists per Decade

A.0.2 Sankey Charts for Topic Relevance

A.0.3 Important of Sentence Transformers

article graphicx

Sentence Transformer	Description	Time Consumption	Advantages	Disadvantages
all-MiniLM-L6-v2	Allround Approach	261 s / 1000 songs	- Quick execution - good results - Greater variety of topics - possibly more detailed insight	- Can produce confusing analyses
all-Mpnet-base-v2	Emotion Detection	100 s (10 songs genre and decade) / 2229 s / 1000 songs	Compromise between focus and topic variety	Not as detailed as MiniLM and compact as LABSE
paraphrase-Mpnet-base-v2	456	208 s (10 songs genre and decade)	- Quick execution - more focused topics, fewer topics	- Fewer results - long execution time
LaBSE	456	90 s (10 songs genre and decade)	- Quick execution - Very compact, ideal for quick analyses	- Variety of topics greatly reduced

Table 1: Comparison between different Sentence Transformers

A.0.4 Important topics over time

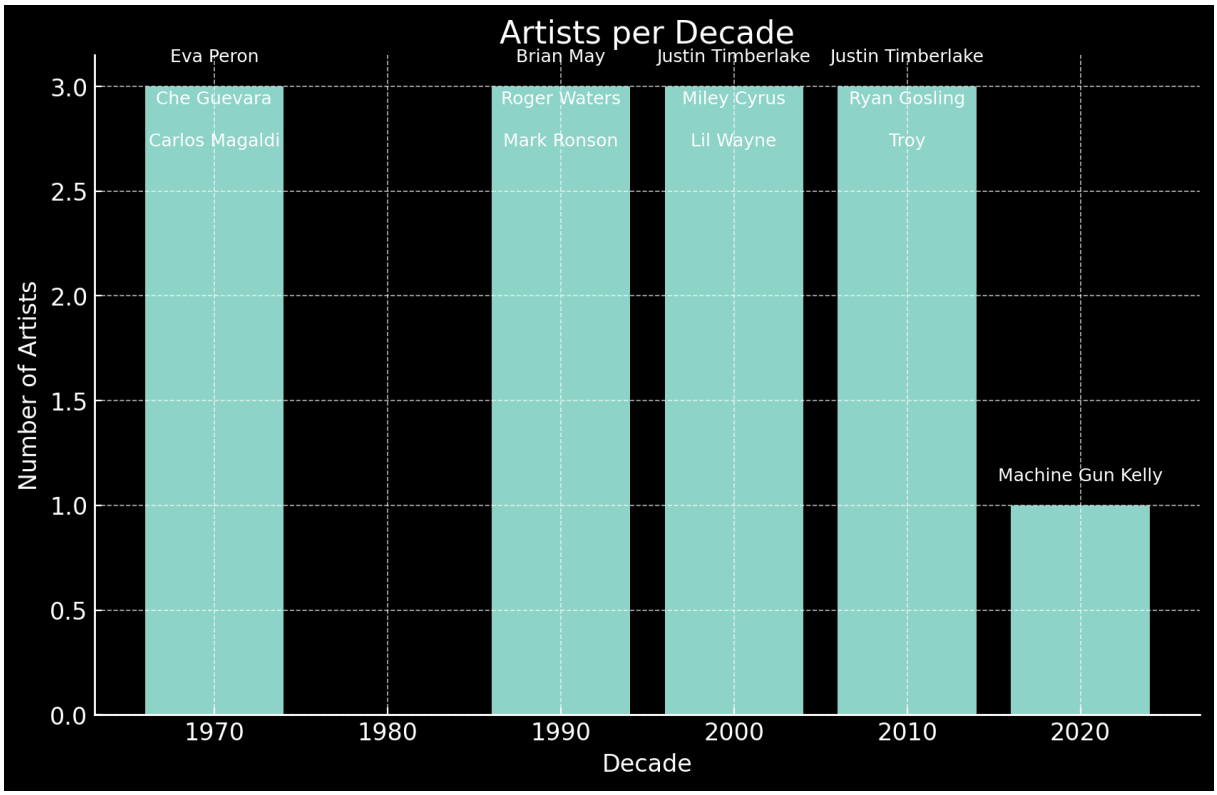
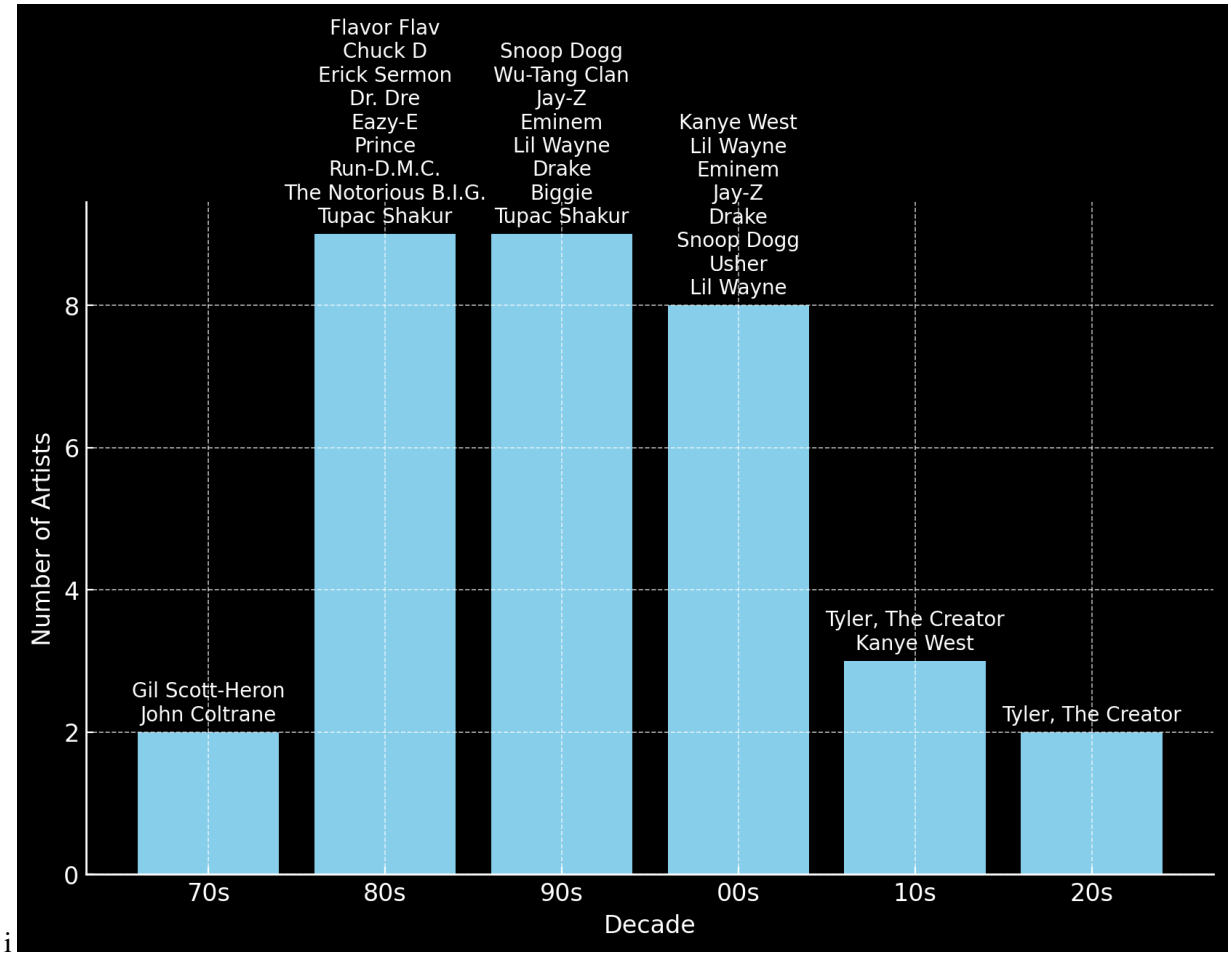


Figure 1: Number of artists per decade using BERTopic POP



i

Figure 2: Number of artists per decade using BERTopic RAP

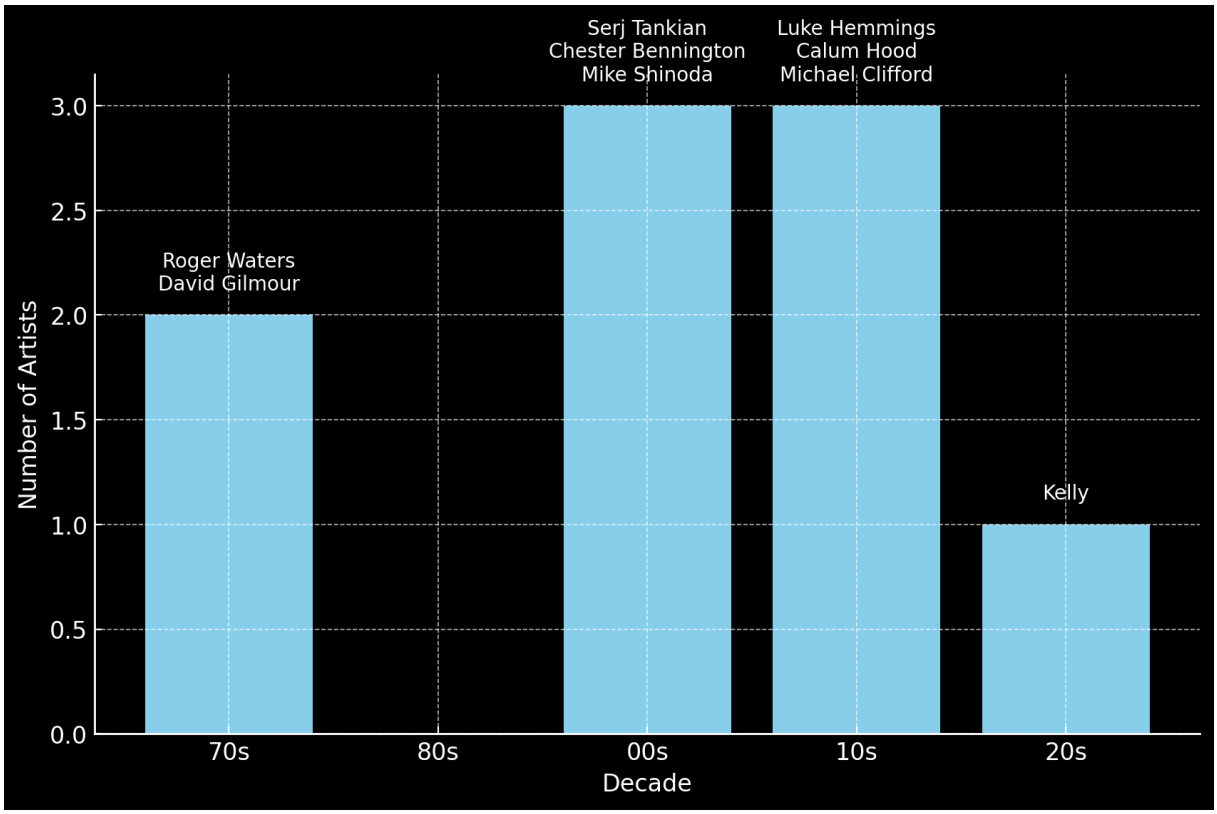


Figure 3: Number of artists per decade using BERTopic ROCK

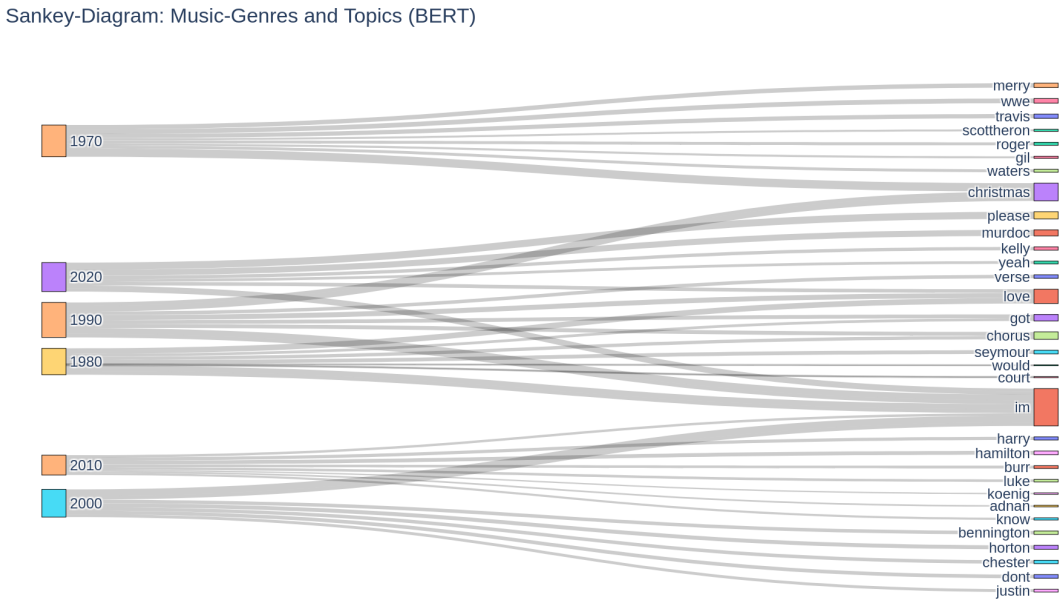


Figure 4: Sankey Diagram music genres and Topics using BERTopic

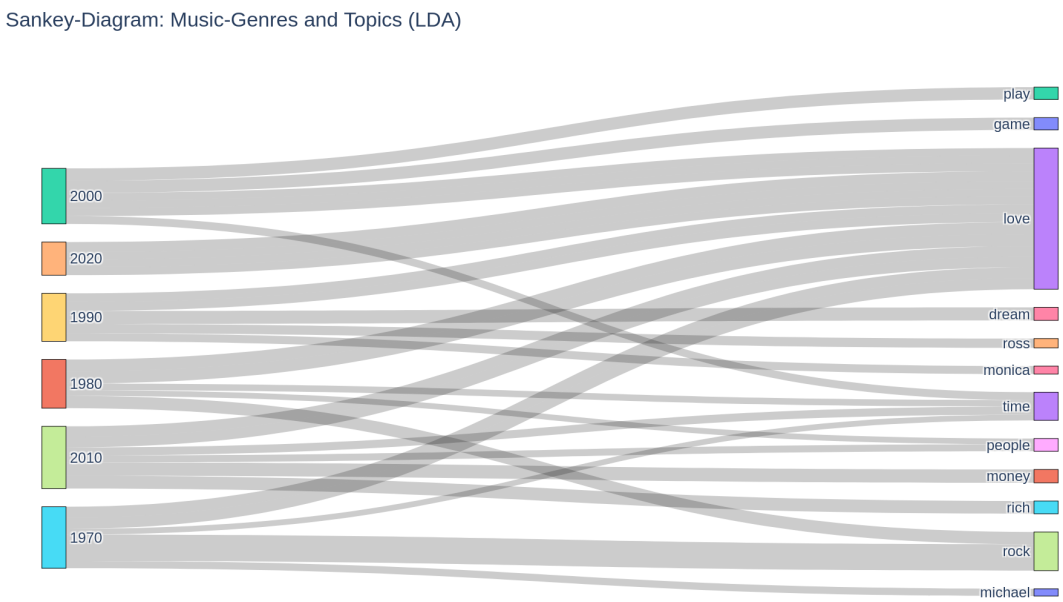


Figure 5: Sankey Diagram music genres and Topics using LDA

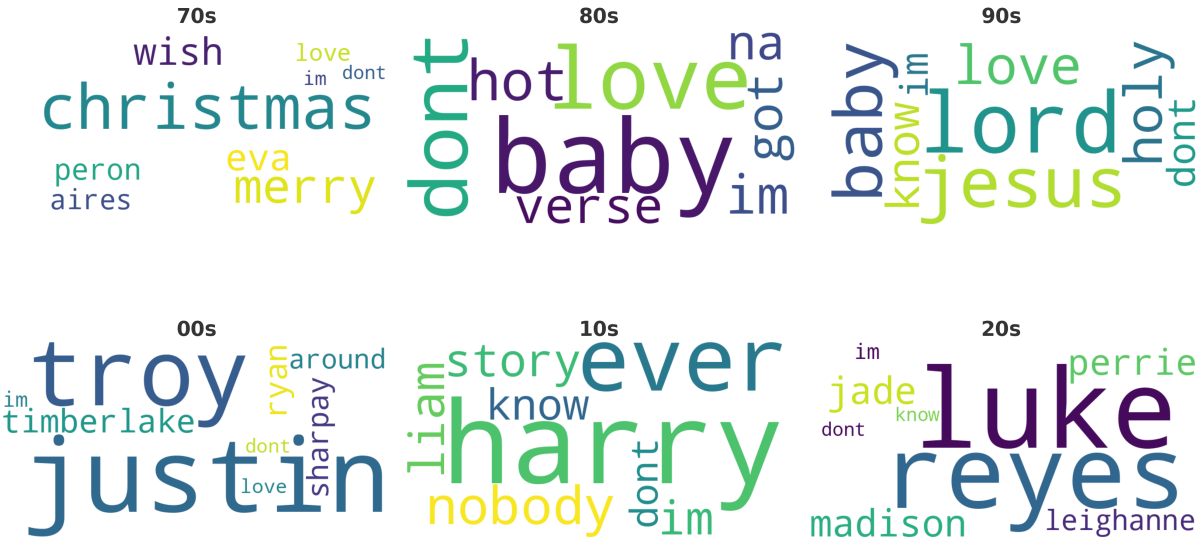


Figure 6: WordCloud important Topics in POP over time, 1970 - 2025

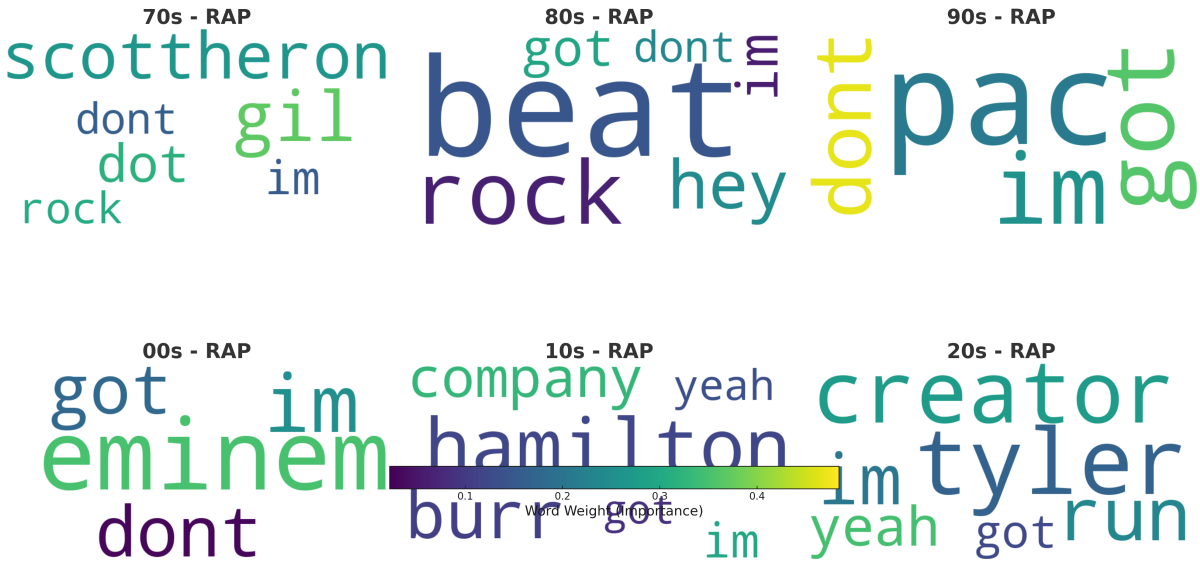


Figure 7: WordCloud important Topics in RAP over time, 1970 - 2025

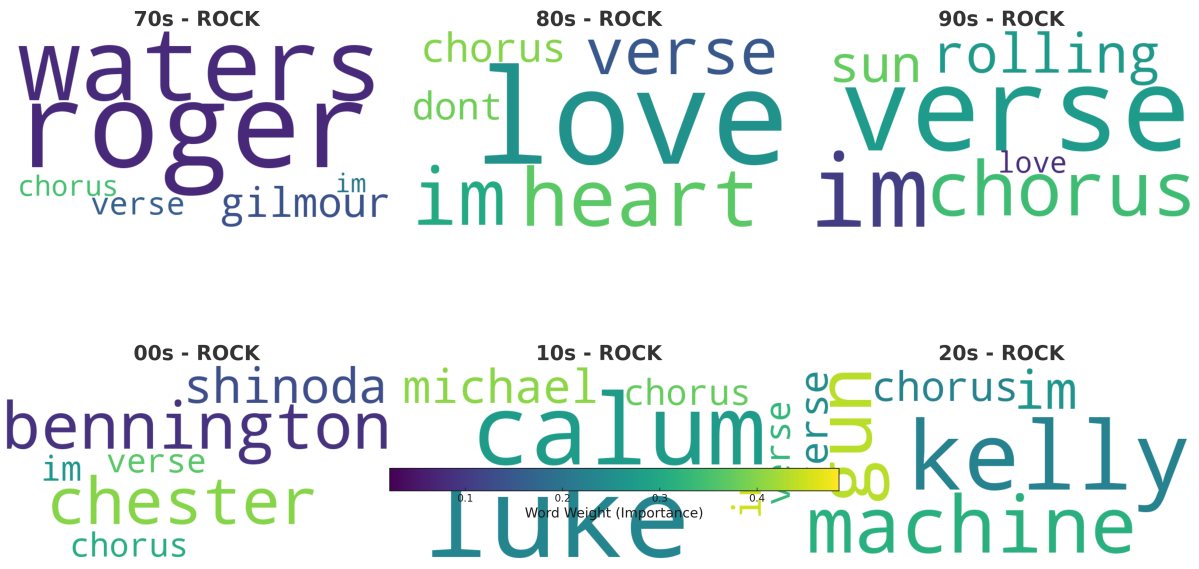


Figure 8: WordCloud important Topics in ROCK over time, 1970 - 2025