INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Sentiment Analysis and Topic Modeling of Portuguese and Brazilian Song Lyrics through the years

Inês Mariana da Trindade D'Alva

Master in Data Science

Supervisor:
Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
Doctor Pedro de Paula Nogueira Ramos, Full Professor,
Iscte – Instituto Universitário de Lisboa

October, 2023

iscte BUSINESS SCHOOL

iscte TECNOLOGIAS E ARQUITETURA

Department of Quantitative Methods for Management and Economics
Department of Information Science and Technology

Sentiment Analysis and Topic Modeling of Portuguese and Brazilian Song Lyrics through the years

Inês Mariana da Trindade D'Alva

Master in Data Science

Supervisor:
Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
Doctor Pedro de Paula Nogueira Ramos, Full Professor,
Iscte – Instituto Universitário de Lisboa

October, 2023

To my mother for being an inspiration
and making me who I am today

Acknowledgment

This research presents an opportunity for me to meet my academic path with my great passion for music. I am really grateful to have had the opportunity to do this type of study, however this was a challenging task and I would like to thank all those who supported me and gave me motivation to complete this.

Starting with expressing my gratitude to have worked with my supervisors Professor Ricardo Ribeiro and Professor Pedro Ramos and I appreciate all the guidance and directions given from them.

I would also like to take this moment to thank my mother for her unconditional support in all the phases in my life, and for being my true hero and inspiration, from her actions to her motivational words, always making sure I had everything needed to be able to complete this chapter of my life.

To my sister and my father, always present throughout this project, a massive thank you, for the deep support and attention expressed, that help me get this through.

And last but not least, I am grateful for the support provided by my boyfriend, Vasco, who would always cheer me up in stressful times, keep me company during late nights of work, be a shoulder for the most challenging times and helping me believe in myself and in my ability to reach my goals.

Resumo

As letras de uma música são uma rica fonte de informação, entre os diversos componentes no contexto musical. Com a sua identidade distinta e elementos narrativos, as letras têm o poder de transmitir mensagens profundas, com as emoções e os sentimentos retratados e os temas abordados. Ao longo do tempo, esses componentes líricos evoluíram, refletindo as mudanças nas dinâmicas da sociedade.

Esta dissertação tem como objetivo estudar essas mudanças de sentimentos e tópicos no cenário nacional de Portugal e Brasil, abrangindo desde a década de 1960 até a década de 2020. Para alcançar estes objetivos, utilizamos uma abordagem baseada em léxico para análise de sentimentos e empregamos BERTopic e LDA para o modelo de tópicos.

Os resultados das nossas pesquisas revelam um contraste emocional entre os dois países. As canções brasileiras predominantemente exalam positividade e sentimentos motivadores, enquanto que as canções portuguesas frequentemente carregam um tom de negatividade. Os tópicos extraídos das letras frequentemente se alinham com as experiências históricas e sociais de cada nação. No entanto, algumas instâncias mostram uma desconexão, onde as letras não refletem com precisão os períodos desafiadores, em termos de tópicos ou polaridades de sentimento. Isso sugere que os letristas podem usar as suas criações musicais como uma forma de escapar à realidade.

Palavras Chave: *Letras de Música*, *Análise de Sentimento*, *Modelação de Tópicos*, *Text Mining*

Abstract

Music lyrics are a rich source of information, within the various components in the musical context. With their distinctive identity and narrative elements, lyrics have the power to convey profound messages, with the emotions and sentiments they portray and the themes addressed. Over time, these lyrical components have evolved, mirroring the changing dynamics of society.

This dissertation aims to study these sentiment and topic changes in the national scope of Portugal and Brazil, spanning from the 1960s to the 2020s. To achieve this, we employ a lexicon-based approach for sentiment analysis and utilize BERTopic and LDA for topic modeling.

The results of our research reveal an emotional contrast between the two countries. Brazilian songs predominantly exude positivity and uplifting sentiments, while Portuguese songs often carry a prevailing undertone of negativity. The extracted topics from the lyrics frequently align with each nation's historical and societal experiences. However, some instances show a disconnect, where lyrics do not accurately mirror challenging periods in terms of topics or sentiment polarities. This suggests that lyricists may employ their musical creations as a form of escape from reality.

KEYWORDS: *Music Lyrics*, *Sentiment Analysis*, *Topic Model*, *Text Mining*

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

"Words or Lyrics is the soul of music"
(Ahuja and Sangal, 2018)

Music contains a great flow of information. Melody, rhythm, and all components in-between, are significant for the message it represents. Especially, the song lyrics. Filled with an identity and story-line elements (Laoh et al., 2018), they are not just words, they are a representation of their writers' emotions, opinions, and thoughts, which are turned into characters and letters (J. Choi et al., 2018), being also a representation of a respective songwriter's style and signature characteristics (Rosebaugh and Shamir, 2022).

As positive and virtuous emotions are put into songs, song lyrics may also contain negativity, and even demonstrate signs of social biases. It has been concluded that there is a correlation between human and lyrics bias, therefore, music lyrics may also portray a societal trend (Barman et al., 2019). Over the years, there is also a change in these components, in the lyrics, consequently, these changes may be perceived as a reflection of society, and they are a good opportunity to study the dynamics of cultural and social changes (Brand et al., 2019) and the narrative of a population, on a psychological level, as stated by DeWall et al. (2011).

As the sentiment they transmit flows, there is also a variation of the topics addressed, that demonstrate a flux of political and social ideals (Napier and Shamir, 2018). Thus, an analysis of the song lyrics can play a crucial part when characterizing and describing financial, cultural, gender, and values evolution (Vieira et al., 2019).

How can that be achieved?

As a computational approach to extract this type of information, there is one of the fields of Natural Language Processing, that is Sentiment Analysis (Devika et al., 2016). Sentiment Analysis, also known as "subject analysis", "opinion mining", or "appraisal extraction" is a field of research that studies subjective elements from words, sentences, phrases, and expressions, providing a polarity classification to determine if the object is "positive" or "negative" (Mejova, 2009). As Nath and Phani put it: "The way something is thought about, in terms of some idea or thought or a feeling, is called the sentiment or opinion. Analysis of this sentiment expressed by humans is called the sentiment analysis or opinion mining" (Nath and Phani, 2021). This approach can be used for multiple types of text objects, however, most of the current literature uses this on product or movie reviews, news, political discussions, and user comments on the web (Mejova, 2009).

Applying this to song analysis is a challenging and complex task, as it requires a thorough selection of emotional features that would provide the best results (Kamalnathan et al., 2019). The fact that music lyrics are, usually, poetic, filled with expressive resources, turns the analysis of these verses harder to achieve (Laoh et al., 2018).

To study the topics addressed in a song, there is the Topic Modeling area. This is "a statistical technique for revealing the underlying semantic structure in large collection of documents" (Kherwa and Bansal, 2020). For these activities, Latent Dirichlet Allocation (LDA) (D. Blei et al., 2001) and BERT models (Devlin et al., 2018) have been used, amongst others.

As part of Natural Language Processing, other techniques have been emerging, which is the case of Transformer-Based models, that have shown to outperform previous algorithms (Agrawal et al., 2021).

This dissertation embarks on a journey to unlock the sentiments and themes concealed within the intricate fabric of song lyrics. It leverages the power of Sentiment Analysis and Topic Modeling, guided by the rich landscape of NLP methodologies, models and techniques. In doing so, we aim to shed light on the emotional and thematic dimensions of music lyrics from Portugal and Brazil over the decades from the 60's to the 2020's, unraveling their complexities, and contributing to the ever-evolving understanding of this expressive art form.

This thesis follows the CRoss-Industry Standard Process for Data Mining (CRISP-DM) methodology. This 6-steps model, developed by Daimler Chrysler and Daimler-Benz (Shafique and Qaiser, 2014), has been proven to be more complete when compared to Sample, Explore, Modify, Model, Assess (SEMMA) and simpler when compared to Knowledge Discovery Databases (KDD), as it incorporates all the phases included in KDD model, consolidating a few of them, to reduce the number of steps (Azevedo and Santos, 2008). This it is described by the following phases, as seen in Figure 1:

(1) *Business Understanding*
(2) *Data Understanding*
(3) *Data Preparation*
(4) *Modeling*
(5) *Evaluation*
(6) *Deployment*

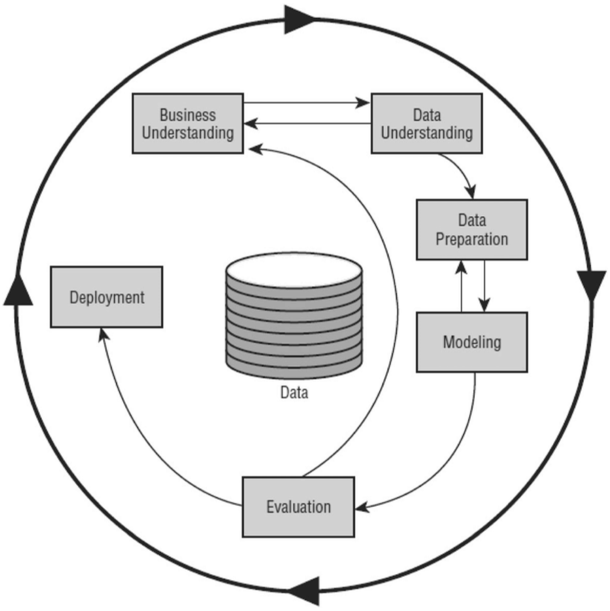


Figure 1. Phases of CRISP-DM (source: Chapman et al., 2000)

This model initiates with understanding the scope of the problem and the questions that the study will answer, so a goal can be defined and the objectives are clear. Once the theoretical problem is well identified, the specific technical goal also needs to be specified, so a good assessment of the tools and necessary techniques to achieve it can be planned accordingly. Moving to Data Understanding, this phase consists of collecting the needed data for the activity, its cleaning processes, and verification of its quality. This is where a full and detailed knowledge of the data is gathered, with exploratory and distribution analysis of the data and its visualization. This is followed by Data Preparation, where the data is formatted and restructured to be used by the models. In the Modeling phase, the models are selected with the appropriate parameter settings. As seen in the above figure, there is an inter-flow between Data Preparation and Modeling, demonstrating that depending on the model applied, different types of preparation tasks may be executed. At this point, with the final model or models, we move to the Evaluation phase, to analyze the model results based on the questions defined in the Data Understanding phase, where it is concluded if the pre-defined goal is achieved. After this is complete, the last phase is Deployment, where the models are planned to be deployed in the business, considering monitoring and maintenance activities. Since this is an academic study, this last step will be skipped (Chapman et al., 2000).

CHAPTER 2

State of the Art

2.1. The Importance of Song Lyrics

Besides audio features, a song can be an aggregation of multiple dimensions, each with its characteristics. From rhythm and melody to the lyrics, one may interpret a song in diverse ways (Souza and Café, 2018). The lyrics of a song are very important, as they contribute to the semantics and harmonics of a song in a different way than the melodies (Besson et al., 1998). A song lyric conveys information about the sentiment, emotion, topics, and themes of the song (Oudenne et al., 2010), which magnifies the relevance of its study.

In these words, we find a reflection of social issues and topics from the ambient they are written in, thus they are a representation of the society (Betti et al., 2023). Davis (1985, as cited in Betti et al., 2023), states that song lyrics "are more than mere mirrors of society; they are a potent force in the shaping of it". As time goes by, the lyrics also go through changes, showing a difference in the emotions and the trending topics through the decades (Frith et al., 2001, as cited in Napier and Shamir, 2018), thus a study of these changes may help in understanding a societal evolution (Brand et al., 2019).

Sérgio Godinho, one of the most influential and respected figures in Portuguese music (as extensively described in Section 3.2, as he is included in the palette of our selected composers), answered to us about how the societal context impacted his songwriting, he stated that "all surrounding environment had an influence [...] and all that, evidently, shapes a social conscious, also political", referring to the event of May 68, that forcibly impacted the artist while his exile in Paris, also mentioning the impact of not being able to return to his homeland. Regarding the influence of others songwriters, Godinho highlights José Mário Branco and Luís Cília affiriming that "the exchange of experiences was determining" when forming his own concept of a song, being also inspired by José (Zeca) Afonso to revamp his work, making him understand that he "could write in Portuguese in another way". On an international perspective, he refers to Chico Buarque, Caetano Veloso, The Beatles, The Rolling Stones, with attention to Bob Dylan.
Regarding the topics addressed in his songs and their shift throughout his career, the

lyricist affirmed that his songs are diverse:

"I believe I address various themes [...]: some more related the everyday, others one may consider to be about Love [...] one of the central topics of my songs, and is an open word, because then, in each song, is treated differently, in a societal content or societal criticism, or even exposing social and political situations that always exist in the society"

The musician admits that his songwriting may go through changes, as he repeats his schemes, noting that these changes come with life experiences, ending his testimony stating that "very often, something happens that triggers an idea for a song."

2.2. Text Mining and Natural Language Processing

Data Mining is a field of Data Science, that consists of the process of extracting information, knowledge, useful insight and patterns from a given collection of data. As a variation of that, Text Mining refers to when the given data is composed of natural language text instead of a structured database. Stating this, text mining is a computational process of analyzing different sets of written data to retrieve information from it (Hearst, 2003). Text classification, clustering, entity relation and event extraction are examples of the type of insights that can be retrieved from the text data (Kao and Poteet, 2007). This approach is commonly applied to texts containing either "factual information or opinions" (Witten, 2004).

In this context, Natural Language Processing (NLP) leverages the syntactic (order and relative positioning in a sentence) and semantic (word meaning) scope (Nassirtoussi et al., 2014) and underlying patterns to enrich the algorithms (of Text Classification, Sentiment Analysis, Topic Modeling, Information Extraction), enabling a deeper exploration compared to traditional Text Mining techniques, that rely only on the word itself (its frequencies, as an example) (Dreisbach et al., 2019). Thus, NLP results in a more meaningful understanding and representation of the unstructured text, considering figures of speech, linguistic concepts, grammar, dependency relations, and ambiguities, achievable by counting with a range of knowledgeable characterization, namely word lexicons with grammatical properties (Kao and Poteet, 2007).

Multiple tasks can be completed by using NLP, such as the following: "automatic summarization, translation, named entity recognition, relationship extraction, speech recognition, part-of-speech tagging and parsing, topic segmentation, and sentiment analysis" (Costa, 2017).

As this study focuses on the two lasts mentioned above, we will dive into them in the following subsections.

2.2.1. Sentiment Analysis

Also known as Opinion Mining, Sentiment Analysis consists in the identification and categorization between positive and negative polarities in a free and unstructured text (Nasukawa and Yi, 2003).
The fundamental application of this task is the extraction of people's opinions, as emphasized by Nielsen (2011) (cited in Rajput, 2020), the words within a text convey the writer's positive or negative perspective on the subject. However, when completing this task it is essential to consider the culture and context implicated, as these factors add a layer of complexity to this computational domain (Rajput, 2020).
Even though this process may also include the identification of the polarity of expressions and their relationship to a certain subject, it is the sentiment recognition that has been seen applied to most of the related work in this field (Nasukawa and Yi, 2003).
Commonly, this is seen applied in extracting opinion from reviews, however, regarding applying sentiment analysis to song lyrics, this has been considered by Oudenne et al. (2010) to be a harder and more challenging task to accomplish due to the following reasons:

(1) the songs may include a mix of sentiments, often initiating in a lower or negative tone and emotion, but ending in an uplifting and expecting mood (and vice-versa);
(2) a song may not include many subjectivity and opinion-related lexicon, with the use of nouns that are not associated with either a positive or negative emotion;
(3) it is also usual to see in a song lyric a positive emotion about a negative event (and vice-versa), which also introduces a mix and potential confusion of well-defined sentiment.

In Madhoushi, Hamdan, and Zainudin's work (2015), we find the most frequently used techniques of Sentiment Analysis in other studies. In this study, the authors consider both Opinion Mining and Topic Modeling as Sentiment Analysis approaches.

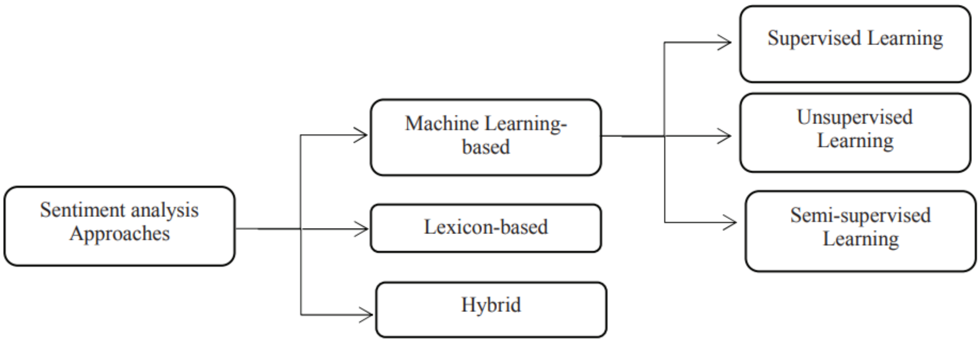


FIGURE 2. Sentiment analysis approaches and techniques (source: Madhoushi et al., 2015)

Dividing the Machine Learning-based models into Supervised, Unsupervised, and Semi-supervised learning types of models, Naïve Bayes, Support Vector Machine (SVM) are the most used supervised models, followed by K-Nearest Neighbors (KNN) (Madhoushi et al., 2015). Supervised learning models are limited in the high sensitivity their results will be based on the training data, showing poor results when the training data does not have a large volume or is biased.

For Unsupervised learning models, these authors emphasize Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (pLSA). The challenge for unsupervised learning is the need for a large quantity of data, so its training is accurate. Semi-supervised learning (SSL) models learn from both labeled and unlabeled data, overcoming the limitation of each type of learning. Self-training, generative models, co-training, multi-view learning, and graph-based methods are the techniques most frequently used when studies apply SSL.

For lexicon-based approaches, Madhoushi et al. segregate this into two types of methods: a dictionary-based approach, which is based on gathering from a dictionary synonyms and antonyms of the given emotion words from the given text; and a corpus-based approach which counts on a corpus to understand the opinion words considering the semantic context (Madhoushi et al., 2015).

According to Mehta and Pandya (2020), amongst the Sentiment Analysis studies, SVM, Naïve Bayes, and Neural Networks have the highest accuracy, being considered the main methods, but also lexicon-based methods have been showing very effective results.

2.2.2. Topic Modeling

As the number of research documents, books, and articles increased, the need to automatically read and classify their underlying subject became highly relevant. To accomplish this, Topic Modeling was developed, as another task part of Text Mining. An

unsupervised mathematical and computational model which primary objective is to discover and reveal themes (Kherwa and Bansal, 2020), extract topics, subjects, and content present in a large set of documents of different types (structured and unstructured), by identifying patterns and clustering of the words (Churchill and Singh, 2022).
Multiple types of models are used in Topic Modeling, but the researchers mostly use Latent semantic analysis, Non-Negative Matrix Factorization, Probabilistic Latent semantic analysis, and Latent Dirichlet Allocation (Kherwa and Bansal, 2020). We may highlight the latest, a Bayesian model, "in which each item of a collection is modeled as a finite mixture over an underlying set of topics" (D. M. Blei et al., 2003). These models have revealed faulty results when dealing with documents that change over a long period of time, when analyzing the differences in the same subject, therefore Blei and Lafferty designed a Dynamic Topic Model (Vayansky and Kumar, 2020) to analyze the evolution of a specific topic (D. M. Blei and Lafferty, 2006). Dynamic Topic Modeling (DTM) consists in a set of techniques applied to textual data to analysze how the same topic changes and evolve through a period of time, and how can be represented in the different times, by incorporating a BERTopic model, which provides capabilities to calculate topic representations at each time step without the need of multiple reruns. Its framework consists in initially creating a generic model, fitting the data without a temporal component, which is later calculates the c-TF-IDF (contextual Term Frequency-Inverse Document Frequency) representation in each time point (Grootendorst, 2021).

2.3. Systematic Literature Review

To find similar research done previously in this field, a Systematic Literature Review was conducted.

The initial approach consisted of a preliminary search[1] about the topic described in the above section, on platforms as Paper Digest[2] and Google Scholar,[3] for a general overview of what are the existing studies. This helped prepare the research question this dissertation should answer as well as the terms to use in the formal search query.

Being one of the largest databases of papers, articles, journals, and conference proceedings, Scopus[4] offers as a free tool, a great collection of enriched scientific articles, extending to, approximately, new 3 million items added every year (Baas et al., 2020). It was the chosen database for this review research to ensure an official and quality excellence of the studies presented, where the following search query was used, searching for Article title, Abstract and Keywords:

[1] This search initiated on November 2022, with the final conclusions on June 2023.

[2] https://www.paper-digest.com/ (Accessed: June 2023)

[3] https://scholar.google.com/ (Accessed: June 2023)

[4] https://www.scopus.com/home.uri (Accessed: June 2023)

lyrics AND (song OR music) AND ("sentiment analysis" OR "topic modeling" OR "machine learning" OR nlp OR "natural language processing")

This search resulted in 225 hits. Using one of the functionalities of Scopus, it was verified a significant increase in the literature in the year 2018, as seen in Figure 3.

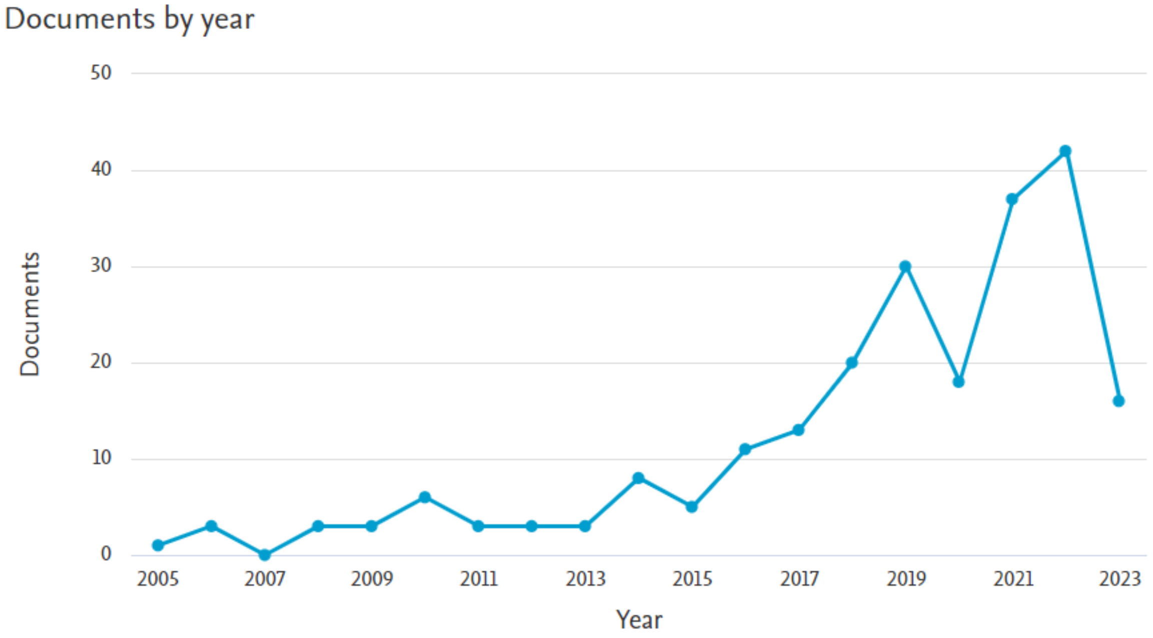


FIGURE 3. Research documents distribution per year in Scopus results (source: Scopus)

The following factors were used as exclusion criteria:

- Algorithm/Model Creation – articles that focused on developing a new algorithm or model
- Consideration of audio features – any article that used audio features (audio signals, acoustics parameters, voice audio) as input for the sentiment analysis
- Conference proceedings
- Corpus/Dataset creation – articles which objectives where to develop a corpus
- Duplicates articles
- Genre/Artist classification
- Inaccessible[5]
- Lyrics, titles and melodies generators
- Not in English or Portuguese language
- Literature review articles
- Other miscellaneous documents[6]
- Recommendation system

[5] Considering any researches that are not accessible in any open-sourced platforms.

[6] Any other document doing other types of studies such as music/lyrics segmentation, lexicon analysis, word similarity and lyrics annotation or not song lyrics related.

As seen in Figure 4, this exclusion resulted in 45 articles classified as *Sentiment Analysis* (28), *Topic Modeling* (12) and *Both* (5), that were considered and analyzed for the present state-of-art. From each, we captured the type of processing method applied, the algorithms or techniques used, the key findings and the target location or region as the main foundation for this analysis.

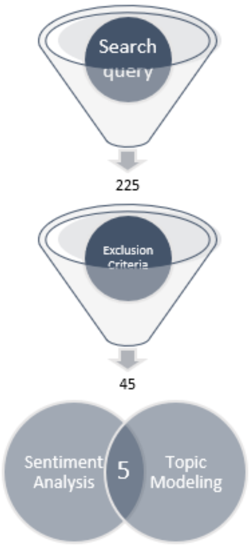


FIGURE 4. Process Workflow

2.3.1. Article Labels

As seen in Table 1, after labeling the 225 articles resulted in the Scopus search, we identified a significant amount of studies (38) that consider not only music lyrics but also audio features as input to develop a sentiment analysis work. In fact, various pieces of research have demonstrated that certain audio features play a crucial part in detecting emotions in music, especially when combined with lyrical features, as studied by Sharma et al. (2020).

Article Label	No. of articles
Algorithm/Model Creation	2
Audio Features (Multi-modal)	38
Conference proceedings	5
Corpus/Dataset creation	6
Duplicates	3
Genre classification	10
Inaccessible	3
Lyrics, titles and melodies generators	28
Not in English or Portuguese	3
Literature review articles	3
Other miscellaneous documents	71
Recommendation System	8
Sentiment Analysis (SA)	28
Topic Modeling (TM)	12
SA + TM	5

TABLE 1. Article Labels

As seen in Table 1, for the 45 considered articles, 12 performed a topic modeling activity, while 28 aimed to detect emotions and mood by analyzing the song lyrics, noting that 5 of the same performed both analyses.

2.3.2. Target Languages

Given that Natural Language Processing algorithms are highly dependent on the languages of the texts and that, currently, its "automatic recognition poses a challenging problem for NLP applications" (Amin et al., 2021), it is relevant to extract the languages for which there is existing literature.

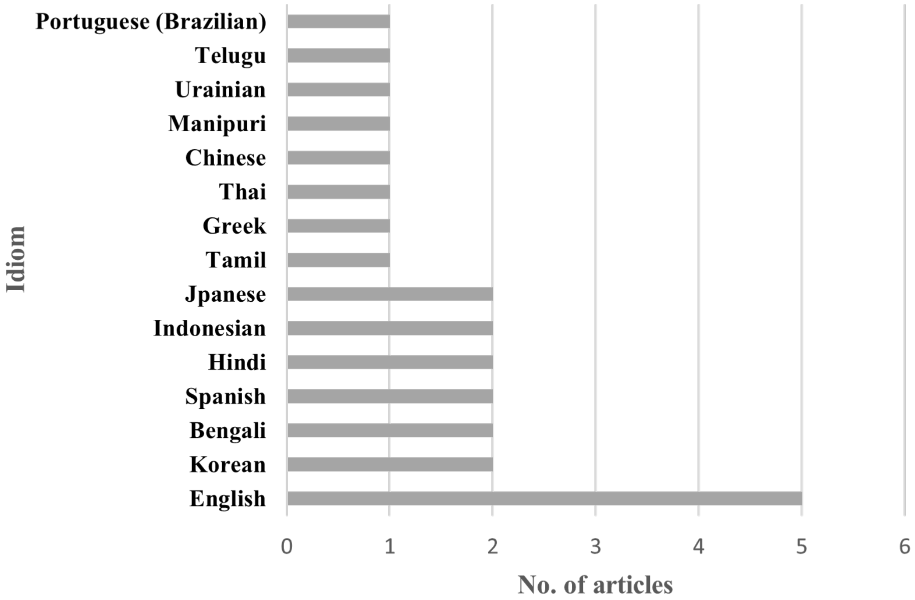


Figure 5. No. of articles per region

Considering only the articles that specify the target language, Figure 5 shows how they are distributed amongst the target location (and corresponding language). It is notable the dominance of languages from Asia/Middle East, such as Hindi, Bengali, and Telugu. Other articles, not included in Figure 5, consider international popular music (from Billboard Top 100), therefore in the English language, in its majority.

2.3.3. Methods and Algorithms

Through the various types of article labels, the methods and algorithms applied in the studies varied accordingly.

The majority of the considered articles performed classification tasks, where Naïve Bayes algorithms were highlighted, used in 12 articles, achieving the best results, when compared to others, as shown by Dang and Shirai (2009), Rajasekar and Geetha (2022), Nath and Phani (2021) and outperforming Support Vector Machine (SVM) (Aziz, Bijaksana, et al., 2019). SVM models are also recurrently used in researches as Malheiro et al., 2016 and Ahuja and Sangal, 2018, presenting greater results in a study approach by Kim and Kwon (2011). More recent researches are focusing on Neural Network models for the sentiment analysis tasks as done by Lee et al. (2018), Laokok and Khonthapagdee (2022) and Rafi-Ur-Rashid et al. (2022). On topic modeling studies, the model most used is Latent Dirichlet allocation, chosen to be applied when the aim is to understand the topics from the songs (Laoh et al., 2018).

Table 2 exhibits a summary of the models and techniques applied in the reviewed literature.

Type	Algorithm / Model / Technique	No. of articles where used
Classifier	Naïve Bayes	12
	Support Vector Machine (SVM)	10
	Logistic Regression	7
	Random Forest	7
	Decision Tree	7
	K-Nearest Neighbor (KNN)	5
	Support Vector Classification	2
	Lexicon-based models	2
	dictionary-based models	1
	Linear Regression	1
	SVM polynomial kernel (PSVM)	1
	Hidden Markov Model (HMM)	1
	Sequential Minimal Optimization (SMO)	1
Neural Network[7]	Convolutional Neural Network (CNN)	2
	Long Short-Term Memory (LSTM)	1
	Bidirectional LSTM (BiLSTM)	1
	Bidirectional Gated Recurrent Unit (BiGRU)	1
Performance Boost	AdaBoost	3
	eXtreme Gradient Boosting (XGBoost)	3
	Light Gradient Boosting Machine (LightGBM)	1
	Gradient Boosting	1
	Bagging	1
	Ridge	1
Similarity	word2vec	3
	Sentiment Vector Space Model (s-VSM)	1
	Graph-based methods	1
Software/Tools	VADER	2
	IBM Tone analyzer	1
	OpinionFinder	1
	WordNet Affect	1
	WEKA	1
	Linguistic Inquiry and Word Count (LIWC)	1
Topic Modeling	Latent Dirichlet Allocation (LDA)	6
	BERT	3
	Structural Topic Modeling (STM)	1
	Heuristic Dirichlet Process (HDP)	1

Table 2. Algorithms, Models and Techniques applied

2.3.4. Summary

In the Appendix section we present a comprehensive summary of the key findings and results extracted from the articles and papers reviewed in the literature review, for those

[7] The presented Neural Networks were also used as Classifiers in the analyzed articles.

that presented most relevant results. This summary is presented in Appendices A and B, providing a clear and concise overview of the various insights, methodologies, and outcomes obtained from each source. By organizing this information systematically, we aim to offer a valuable resource for readers, enabling them to grasp the diverse contributions and perspectives presented in the literature. The tables serve as a reference point for the subsequent chapters of this dissertation, facilitating the discussion, analysis, and synthesis and reference existing research in the field of sentiment analysis and topic modeling applied to song lyrics.

2.4. Major Findings

In alignment with the future end goal that led to this literature review, it is interesting to highlight a few of the studies that present strong similarities with the proposed research.

Starting with Jo and Kim (2023), who conducted a study for which the primary objective was to dive into the emotional motivations of K-pop song lyrics and their intriguing evolution across a span of three decades, from the 1990s through the year of 2019. This ambitious research aimed not only to understand the lyrical sentiments prevalent in this dynamic musical genre but also to discern the nuanced shifts and dynamics that have unfolded over time, with a cultural and societal comparison. It is worth noting that "K-pop has become a truly global phenomenon", with its catchy melodies and captivating performances, has garnered a massive and devoted global following (Romano, 2018). Thus, exploring the emotional resonance of its lyrics is of particular interest, given the genre's significant cultural influence.

This analysis nested on songs that earned popularity making it to the Top 100 charts during the addressed period, considered to convey the sentiments and aspirations of a significant segment of the South Korean populace.

Jo and Kim started by following a path to scrutinize and extract the prevalence of adjectives, emphasising on the ones that indicated underlying sentiment, where the authors introduced the innovative Structural Topic Modeling (STM) approach. This methodological choice was strategic, as it enabled the researchers to traverse beyond mere adjective counts and get into the intrinsic emotion in the terms, besides of offering the capability to extract and visualize the distribution of topics over time (Bai et al., 2021), providing a nuanced and data-driven view of the evolving emotional landscape of K-pop.

This study revealed a consistent increase in the usage of positive adjectives over the three-decade period considered, as the Pearson correlation test conducted conclude that adjectives such as "joyful", "inspiring" and "uplifting", associated with positivity, exhibited a statistically significant increase in frequency. However, the duo also demonstrated adjectives such as "hurtful", "melancholic", and "somber" categorically negative in nature, also displayed a significant increase in frequency over time. Notably, they have encounter a paradox of simultaneous positive and negative emotional expressions in K-pop's lyrical evolution.

On top of what was described above, their study shows that 100% of the extracted positive adjectives and approximately 50% of the positive topics demonstrated an upward trajectory in their usage over time. This phenomenon hints at a concerted effort by K-pop lyricists to infuse their compositions with an increasingly positive and uplifting sentiment, perhaps mirroring the genre's mission to provide a sense of escapism and inspiration to its global audience. However, when looking at the negative emotional landscape, roughly 33% of the negative adjectives and, approximately, 56% of the negative topics exhibited a declining trend, suggesting a subtle shift away from certain expressions of negativity within the lyrical narratives. The authors still salient a message of "a description of an end, or anger", arguing to be related to a motivational message of anger management as opposed to a negative sentiment.

As a possible explanation for the surge in positive sentiment observed over the course of 30 years, Jo and Kim also considered the backdrop of South Korea's economic growth. This remarkable transformation empowered individuals, both in terms of self-realization and increased purchasing power, potentially contributing to an enhanced sense of positivity within the lyrical fabric of K-pop. This economic context provides an essential sociocultural background against which to interpret the emotional landscape of the music.

Thus, these authors contribute by not only offering an understanding of the linguistic shifts but also by providing a window into the broader cultural and societal dynamics that have shaped our understanding of music and its lyrical expressions over the decades, as also studied three years back by Vieira et al. (2019), as we will analyse in the next paragraph.

A deep exploration on the vocabulary dynamics of song lyrics from 1960 to 2009 was performed. This research sought to unravel the potential linguistic shifts, tracking not only the variations in word usage but also the potential transformations in semantic meaning over the course of nearly five decades. To undertake this, the researchers focused on song lyrics sourced from the Billboard Hot 100 chart, which is "the definitive weekly ranking of America's most popular songs" (Whitburn, 1996, as cited by Bradlow and Fader, 2001).

As a metric in this investigation, the concept of *entropy* is used, measuring a word frequency variation over time.

The study proved that certain words exhibited an undeniable affinity for specific decades, almost as if they were emblematic of their respective eras. For instance, the term "Supermodel" displayed high flashes from the 2000s forward, while "Quarrel" made his presence in the '60s, according to the authors' findings. Contrasting with specific terms that reveal a timeless characteristic, with uniformity across the decades, such as "Love" and "Colored".

An intrigue discovery, is a semantic evolution over time, as even when the word remained the same throughout the years, their meanings carried the cultural nuances, as verified for the term “Colored”, that initially had a non-Caucasian individuals connotation, but during the socially charged 1970s, its meaning expanded to spectrum of celestial phenomena, such as the resplendent colors of the sky and the shades of dawn.
With this study, Vieira et al. provide a glimpse into the linguistic shifts but also provides a window into the broader cultural and societal dynamics that have shaped our understanding of music and its lyrical expressions over the decades.

Lastly, let us focus in an extensive investigation conducted by Napier and Shamir (2018), with the objective to unravel the dynamics of emotions within music, from the 1950s to 2016. As done by Jo and Kim, for this research the songs were also collected from the Billboard charts, to gather the most popular tunes of their respective years. The lyrics were collected from sources as MetroLyrics[8] and Genius,[9] ensuring a robust and trustful dataset.
By getting the emotions using the IBM Watson Tone Analyzer tool, at the heart of this research, lays also the question of the relationship between sentiment and time, that the authors decided to respond by using statistical tools. Their analysis of the Pearson correlation coefficient, which results revealed a pattern: a notable increase in the sentiments of Anger, Disgust, and Fear as the years unfolded, with an emerge in the decade of the 1980s. This represented a transition from the more exuberant and joyful sentiments of earlier decades to a realm marked by an augmented presence of complex and intense emotions.
Regarding positive emotions, as the sentiment of Joy, although it had held a prominent position from the 1950s through the 1970s, the subsequent decades witnessed a gradual decline in the prevalence of expressions of this kind within song lyrics. This switch highlights a transition from an era characterized by exuberant and carefree sentiments to one marked by a more diverse emotional palette. Interestingly, the emotion of Sadness, while not dominant in earlier decades, began its ascent during the 1980s, reaching its maximum between 2003 and 2004.
In their analysis, Napier and Shamir (2018) concluded that there was an unmistakable and “clear trend toward a more negative tone in pop music lyrics, with a more significant change around the early 1990s” (Napier and Shamir, 2018). This transition was characterized by an increase in emotions such as Anger, Fear, and Sadness over time. These findings illuminate the dynamic and ever-evolving nature of music as a mirror to the human experience, echoing the shifting tides of emotion and society.

Having these enriched researches as a starting point, it increases the interest in the ambitious goal, as this journey we will be able to expand the range and palette on this area, enlarging for a deeper analysis on songs written in Portuguese.

[8] https://metrolyrics.pro/ (Accessed: June 2023)
[9] https://genius.com/ (Accessed: June 2023)

CHAPTER 3

Methodology

This chapter consists in the explanation of the application of the CRISP-DM methodology, in its various steps.

3.1. Business Understanding: The soul behind song lyrics

The primary objective of this research is to dive into an exploration of the evolution of Portuguese and Brazilian song lyrics spanning from the tumultuous 1960s to the contemporary landscape of the 2020s. This goal is in hand with understanding the intricate narratives embedded within these lyrics and how they correlate with the historical tapestry of each nation. Specifically, we seek to discern whether these lyrical compositions serve as reflective mirrors of their respective countries' histories, deftly articulating the prevailing cultural and societal circumstances of their times.

To untangle the essence of these lyrical works, we apply a multifaceted approach, taking advantage of sophisticated analytical models that encompass sentiment analysis and topic modeling. By using these tools, we aim to not only assess the emotional polarities coursing through the lyrics — whether they are positive or negative — but also to extract the themes beneath the surface of the songs. From a technical perspective, our ambition is to develop meticulously models by successfully executing text mining tasks to boost the models into getting insight of the songs.

Starting with a demanding selection of songwriters who can authentically represent their respective homelands, we will create a dataset with their song lyrics. Subsequently, we develop the sentiment analysis and topic modeling models to dissect the lyrical content. Finally, we uptake an extensive analysis of the results, in a journey that will interlace the findings with the broader historical, cultural and societal transformation that have been happening in both Portugal and Brazil over the decades, to expand the knowledge on the relationship between music, lyrics, and the socio-cultural tapestry of nations. The Figure 6 summarizes the planned tasks to complete what has been described.

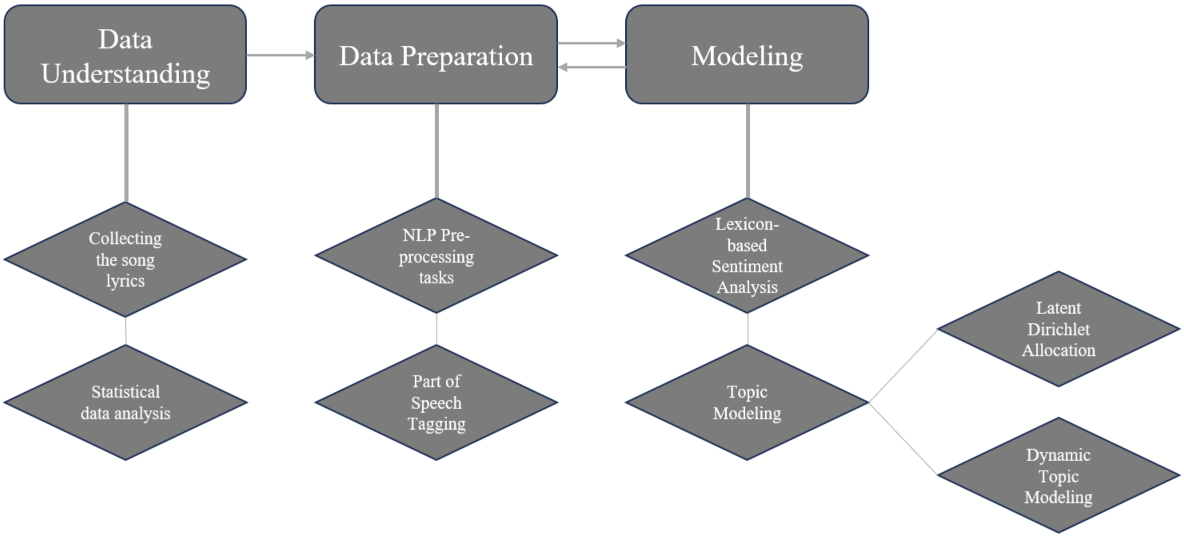


Figure 6. Planned tasks for the project for Data Understanding, Data Preparation and Modeling phases

3.2. Data Understanding: Collecting the lyrics

For this study, a group of seven remarkable lyricists was carefully selected, from both Portugal and Brazil. From each lyricists, we have chosen to include only songs which lyrics were written by themselves. Each of these artists brings a unique blend of age, career longevity, lyrical prowess and excellence to the table, being, in the majority of the songs they sing, its authors and lyricists.

In Portugal, our lineup features the talents of Fausto (74 years old), Jorge Palma (73 years old), José Mário Branco (deceased in 2019 at 77 years of age) and Sérgio Godinho (78 years old). These luminaries have not only stood the test of career time but have also left an outstanding mark on the national music scene. Their lyrical masterpieces have earned them prestigious awards. For instance, José Mário Branco won twice the José Afonso Award, which anually distinguishes portuguese music that have the Portugal culture and history as a foundation. The same award was granted to Jorge Palma in the year of 2002, collecting the Pedro Osório Award (which recognizes Portuguese music and its authors) for his album "Com Todo o Respeito", 10 years later, amongst others. Another example of awards was the Authors Award - Best Album received by Fausto on his album "Em Busca das Montanhas Azuis", and last but not least, Sérgio Godinho's double recognition on the José Afonso Award.
Notably, these gifted artists are leaving a trail in the hearts of their listeners.

On the musical side of Brazil, we welcome the iconic trio of Caetano Veloso (81 years old), Chico Buarque (79 years old) and Gilberto Gil (81 years old). These lyrical personalities have a long career and have consistently delivering songs that resonate deeply with audiences, coming to times where they used their poems as weapons. Their musical presence mirrors pivotal national events, capturing the essence of the people, their emotions, and the prevailing zeitgeist. Also awards collectors, with Caetano Veloso's "Cê" earning him a Latin Grammy Award in the Best Singer-Songwriter Album category in 2007, Chico Buarque's "Tua Cantiga" achieving recognition as Best Song on the Brazilian Music Awards, and recently receiving the Camões Prize, recognizing lyrical work in the Portuguese Language, and Gilberto Gil's with the Grammy Award for Best Contemporary World Music Album in "Eletracústico" in the year 1998, setting the bar as the first in the Federal Insitute of Bahia to receive *honoris causa*[1] recognition, due to his contributions to the Brazilian Culture.

What sets these lyricists apart is not only their ability to interpret the human experience through words but also their talent for creating verses that transcend generations. As prolific songwriters, they have also lent their lyrical talents to other celebrated artists, leaving an indelible mark on the world of music. Through their verses, we attempt to understand the very essence of their nations, their people and their times.

3.2.1. Web Scraping

To extract the lyrics from websites, web scraping methods were utilized. Web Scraping is an automatic, pre-coded, and fast extraction of unstructured data from the web, organizing it into a structured dataset that can be used for analysis (Sirisuriya et al., 2015a).
For this work, the Web Scrapper Google Chrome Extension was used. This tool is based on the SELECTORS on a website. To use this tool, a plan is configured to organize how the website should be navigated and which data to extract, which is later exported as a CSV file (Sirisuriya et al., 2015b). The configuration followed the guidelines described in the Web Scraper website.[2]

[1] Recognition of honor granted by a high prestige university to a personality due to their highly important work, activity or service provided to the country or humanity.

[2] https://webscraper.io/documentation/scraping-a-site (Accessed: August 2023)

3.2.2. PDF Conversion

In addition to the lyrics extracted from websites, a subset of those included in this study was extracted from books and CD's booklets. For this activity, a PDF Conversion was conducted in R with *pdftools*[3] and *tesseract*[4] packages.
The *pdftools* receives a pdf file and rendered its pages into images in a given resolution, which is then passed to *tesseract::ocr*. The resolution (dpi value) used was 600. The *tesseract* package was also used when extracting the lyrics from pages that had more than one column.
This package was not error proofed for the Portuguese language, presenting some errors, especially in words with accents or cedilla marks, resulting in a manual review and cleanup effort.

3.2.3. Song Lyrics Collection

This section is split between the seven songwriters included in our scope, as the lyrics were collected differently for each. After the data gathering from each songwriter, they were joined forming the dataset for this study.

3.2.3.1. *Chico Buarque*

Chico Buarque's lyrics were collected from his official website,[5] via the web scraping method previously explained. The year of the song, its title and its lyrics, were collected from the website using the components highlighted in Appendices C and D.
From the data collected, songs labeled as "Out of Discography" by the website and also those with no lyrics available have been excluded from the scope.

This data gathering resulted in a dataset with the structure present in Table 3.

Variable	Description
year	Year of release
song_title	Title of the song
lyrics	Lyrics of the song

TABLE 3. List of features gathered for Chico Buarque's song lyrics

[3] https://cran.r-project.org/web/packages/pdftools/index.html (Accessed: August 2023)
[4] https://cran.r-project.org/web/packages/tesseract/index.html (Accessed: August 2023)
[5] http://www.chicobuarque.com.br/ (Accessed: March 2023)

3.2.3.2. *Gilberto Gil*

As done for the previous songwriter, the Google Web Scraper tool was used for Gilberto Gil's songs. Using his official website,[6] the lyrics were collected from the songs in each of his album, extracting the components highlighted in Appendices E, F, and G.

From this web scrapping, the data gathering resulted in a dataset with the structure present in Table 4.

Variable	**Description**
album	Title of the album
year	Year of album's release date
song_title	Title of the song
lyrics	Lyrics of the song

Table 4. List of features gathered for Gilberto Gil's song lyrics

After the data collected, songs from Soundtracks, Compilations, and Collaborations albums were excluded.

3.2.3.3. *Caetano Veloso*

For Caetano Veloso's songs, the lyrics had already been pre-extracted from his official website,[7] using also a web scraping method, using the *rvest* package in R.[8]

3.2.3.4. *Fausto*

Given that there is no existing official website of Fausto's, the lyrics were individually collected from his albums of originals, as seen in Table 5.

[6] https://gilbertogil.com.br/ (Accessed: March 2023)

[7] https://www.caetanoveloso.com.br/blog/ (Accessed: March 2023)

[8] https://cran.r-project.org/web/packages/rvest/index.html (Accessed: March 2023)

Album Name	**Release Date**
Fausto	1970
P'ró Que Der e Vier[9]	1974
Um Beco com saída[9]	1975
Madrugada dos Trapeiros	1977
Histórias de Viageiros	1979
Por Este Rio Acima	1982
O despertar dos alquimistas	1985
Para além das cordilheiras	1987
A preto e branco	1988
Crónicas da terra ardente	1994
A Ópera Mágica do Cantor Maldito	2003
Em Busca das Montanhas Azuis	2011

Table 5. Fausto's albums

3.2.3.5. *Jorge Palma*

For Jorge Palma's song lyrics, the data was firstly collected from a book of his original song poems called "Na Terra dos Sonhos" (Palma and Callixto, 2005). This book, released in 2005, and organized by João Carlos Callixto, includes all his song lyrics written by himself up to this date.

From this collection, there were missing songs from his two latest albums: Voo Nocturno (2007) and Com Todo o Respeito (2011). To gather this data, the song lyrics were individually collected from each of these albums booklets.

3.2.3.6. *José Mário Branco*

The list of José Mário Branco's songs written by himself was collected from an archive website,[10] composed by a digital database of José Mário Branco's musical work, including music sheets, albums mapping, and song lyrics.
From the list collected, the available lyrics were manually collected from the same website, resulting in 74 lyrics. The remaining lyrics were gathered from Genius,[11] Letras.Mus,[12] and Observatório da Canção de Protesto.[13]

[9] Due to inability to the access these albums, the lyrics of the songs included in them were collected from the Genius website.

[10] https://arquivojosemariobranco.fcsh.unl.pt/ (Accessed: March 2023)

[11] https://genius.com/ (Accessed: March 2023)

[12] https://www.letras.mus.br/ (Accessed: March 2023)

[13] https://ocprotesto.org/portfolio/jose-mario-branco/ (Accessed: March 2023)

3.2.3.7. *Sérgio Godinho*

Similarly with Jorge Palma's lyrics, Sérgio Godinho's were initially collected from the book "Canções de Sérgio Godinho" (Godinho and Saraiva, 1983), organized as a collaboration from Sérgio Godinho and Arnaldo Saraiva, which consolidates a collection of the artist's song lyrics from 1971 until 1983.

The remaining song lyrics were collected individually from his album of originals, as seen in Table 6.

Album name	**Release Date**
Salão de festas[14]	1984
Na vida real[14]	1986
Aos amores[14]	1989
Tinta Permanente	1993
Domingo no Mundo	1997
Lupa	2000
Ligação Direta	2006
Mútuo Consentimento[14]	2011
Nação Valente	2018

TABLE 6. Sérgio Godinho's albums

3.2.4. Additional variables

After all the data was collected and joined into a dataset, an additional variable *language* was introduced to determine the language in which the song lyric is written, since all of the included artists wrote in other languages different from Portuguese.
For this, the *detect_language*[15] function was used, and all lyrics for which the detected language was not Portuguese were removed. For the lyrics for which the function did not detect a language (N/A), a manual review was done to verify their actual language, including the Portuguese ones for the analysis.

Additional, it was created a function in R, to add a new variable for the lyrics' decade.

[14] Due to inability to the access these albums, the lyrics of the songs included in them were collected from the Genius website.

[15] https://www.rdocumentation.org/packages/aws.comprehend/versions/0.2.1/topics/detect_language (Accessed: March 2023)

3.2.5. Dataset Summary

The final dataset counted on six variables, which description is present in Table 7, consisting in a total of 1308 lyrics, distributed by writer as described in Table 8. The dataset has a majority of Brazilian songs, making a percentage of 61%, with a peak in the song count in the decades of the 1970s and the 1980s.

Variable	Description
country	Respective country of the lyricist (Portugal/Brasil)
lyricist	Lyricist name of the lyrics
year	Year of the song
decade	Decade of the song
song_title	Title of the song
lyrics	Lyrics of the song

TABLE 7. Dataset variables

Lyricist	Count
Caetano Veloso	237
Chico Buarque	309
Fausto	117
Gilberto Gil	257
Jorge Palma	140
José Mário Branco	83
Sérgio Godinho	165

TABLE 8. Number of songs per lyricist

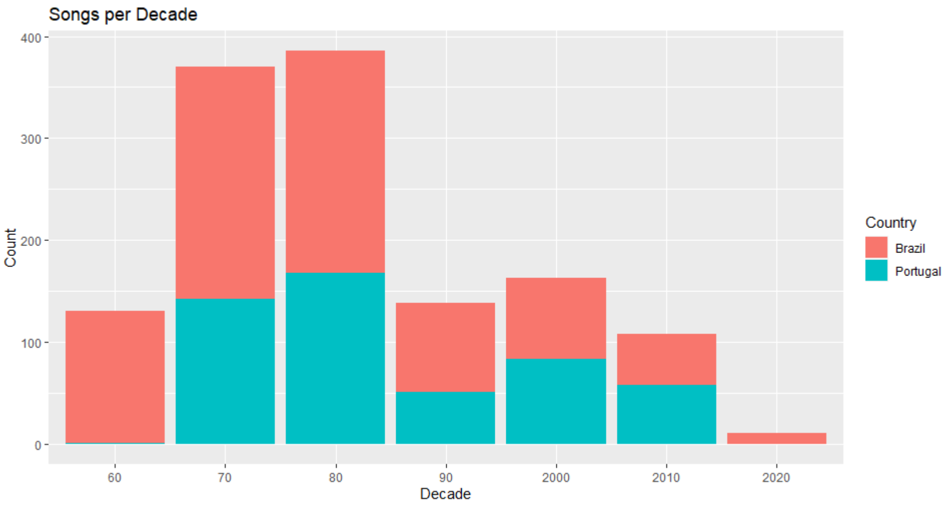


FIGURE 7. Number of Songs per Decade

3.3. Data Preparation: Pre-processing

To prepare the dataset for the topic models, a set of Text Mining pre-processing techniques was conducted. These tasks were performed on Python, using NLTK.[16]

Tokenization

Tokenization is widely recognized as the foundational step in the intricate process of pre-processing for Text Mining. This critical task involves the identification of what are conventionally referred to as *words*, which serve as the fundamental units within a sentence (Habert et al., 1998). While it may appear straightforward on the surface, tokenization is a task that carries a variety of challenges, as astutely pointed out by Webster and Kit. These challenges manifest in the form of diverse delimiters and language-specific nuances (Webster and Kit, 1992), further highlighting the significance of tokenization as the essential basis upon which subsequent text analysis tasks are built.

Stop Words Removal

Words that provide no semantic information to the text are called *stop words.* These are, usually, articles, prepositions, auxiliary verbs, and conjunctions. It has been studied the benefits of removing these types of words, as their contributing information potentially misleads the learning machine (Silva and Ribeiro, 2003). Thus, a custom list of stop words was created, using a Portuguese delivered stop words list, from the NLTK package, and a list of words collected from the dataset, manually considered as *stop words.* This list includes words as interjections such as *ah* or *oh*, contractions of *stop words*, such as *p'ra* or *p'lo*, *nao* instead of *não*, words from other languages (e.g.: *hi* or *you*) that were included in the lyrics, and other words that were found that did not provided no relevant content (e.g.: *tão*, *tanto*, *ainda*, *ora*, *porque*).

When preparing the data for the sentiment analysis, the negation words were kept, so we could handle negation properly.

Non-portuguese lyrics

Although a pre-assessment was done initially to remove any non-portuguese lyrics, after a few iterations of the model, it was found that some non-portuguese lyrics were still included in the dataset. Thus, a new function was created, leveraging the *langdetect*[17] package, to remove these lyrics from our dataset.

[16] Version 3.8.1
[17] Version 1.0.9

This process flagged 10 Brazilian lyrics, that after manual review, only 5 were removed, leaving our final dataset with 1303 lyrics.

Bigrams

As part of the pre-processing tasks, an N-gram analysis was also performed, as this is also a relevant part of Natural Language Processing (NLP). This analysis is based on the identification of a sequence of n-word elements, that commonly appear together (Sidorov et al., 2014), which significantly improve the performance of topic models (Nokel and Loukachevitch, 2015). This is usually used to distinguish the semantic value of a n words combined from when they are used separately (for example "New York" and "Social Media").

For this activity, the *collocations* module[18] from *nltk* was used and only those with an absolute frequency of 8 or above were contemplated. Based on the conducted analysis, between bigrams and trigrams, the data preparation included bigrams identification, as it was verified that the trigrams did not provide relevant information for the study, in the collected dataset.

After the output list of bigrams provided by the model, a manual analysis was done to include only the relevant ones. Finally, in Table 9 is the list of bigrams considered for the topic model.

fim_mundo	centro_comercial	maggoo_polly	banda_militar
beira_mar	direito_humano	bairro_alto	d_sebastião
guiné_bissau	tiro_liro	santo_amaro	bater_porta
quarto_feira	fim_mês	peter_gast[19]	bacalhau_bastar[20]
polly_maggoo[21]	tomar_conta	bomba_relógio	sol_nascer
quente_cabeça	beira_rio	bater_coração	morto_vivo
cabeça_frio	homem_fantasma	edyth_cooper	faz_tempo
sexy_sadie	bandeira_branco	jeca_tatu[22]	raça_humano
fundo_mar	levar_mal	faz_parte	graça_divinar

Table 9. Bigrams considered for the topic models

[18] https://www.nltk.org/_modules/nltk/collocations.html (Accessed: April 2023)

Lemmatization

To create more robust vectorial representations of the lyrics, the lemmatization process was used. This is the process of reducing the word to its base form (also called *lemma*) (Perera and Witte, 2005), by removing its suffixes and prefixes (Kharis et al., 2021). This task can be complex when dealing with verbs in their irregular form.
To complete this task, the lemmatizer[23] from the Spacy[24] package was utilized, with the *pt_core_news_sm* model,[25] which is specifically designed to support the Portuguese language.

Number and Punctuation Removal and Lowercasing

Furthermore, the lyrics underwent a process of eliminating all numerical values and punctuation marks, while also being uniformly converted to lowercase.

In Table 10 it is visible an example of a sentence that went through the applied process, demonstrating *lemmatization*, *lowercasing*, *stop words removal* and *bigrams recognition.*

Original lyrics	Lyrics after pre-process for Topic Modeling	Lyrics after pre-process for Sentiment Analysis
Vem		
entra na minha casa	entrar casa	entrar casa
come a carne abre as gavetas	comer carne abrir gaveta	comer carne abrir gaveta
leva a roupa e as camisas e os postais	levar roupa camisa postal	levar roupa camisa postal
vai	fim_mundo	fim mundo
dizer ao fim do mundo	palavra escorrer	palavra escorrer
as palavras que escorreram	garganta gritar	garganta gritar
na garganta dos que gritaram demais		

Table 10. Portion of lyrics after Text Mining processing techniques

[19] A German music writer (1854-1918)

[20] Reference to the Portuguese popular expression “para quem é, bacalhau basta”, indicating that for someone who is not very important, anything is good enough.

[21] An American supermodel character from the 1966 movie “Who Are You, Polly Maggoo?”, directed by William Klein

[22] Character created by Monteiro Lobato, in a representation of farm worker from São Paulo, Brazil, subjugated to economic and social precarious conditions.

[23] https://spacy.io/api/lemmatizer (Accessed: April 2023)

[24] Version 3.6.1

[25] https://spacy.io/models/pt (Accessed: April 2023); Version 3.6.0

Part of Speech

Recognized since a long time ago, *part of speech* is the distinction of word classes, based on the grammatical distribution, between one of the following: noun, pronoun, verb, adverb, adjective, participle, article and conjunction (Voutilainen, 2003). The *tagging* is the automatic assignment of one of the classes mentioned to all tokens in a given sentence (Voutilainen, 2003). This task is very common in the NLP studies (Martinez, 2012).
In our project, this was used to detect the Nouns, Verbs and Adjectives, as these are classes considered good indicators in the analysis of opinion, topic and sentiment (Costa, 2017), leveraging the the Part-of-speech tagging[26] from the Spacy package, extracting the Universal POS tags of "NOUN", "VERB" and "ADJ", respectively.

Keywords

Another relevant task in the realm of topic detection is keyword extraction. This process entails extracting pertinent information from the text data, consolidating the "key topics", and acquiring the associated "keywords" (Gupta, 2017).
Here, the Rake-nltk[27] package was used. In Figure 8 it is visible a WordCloud of the keywords extracted from the entire dataset, where we can verify that Love, Life and Power ("amor", "vida" and "poder", respectively, in Portuguese) are relevant themes in the songs.

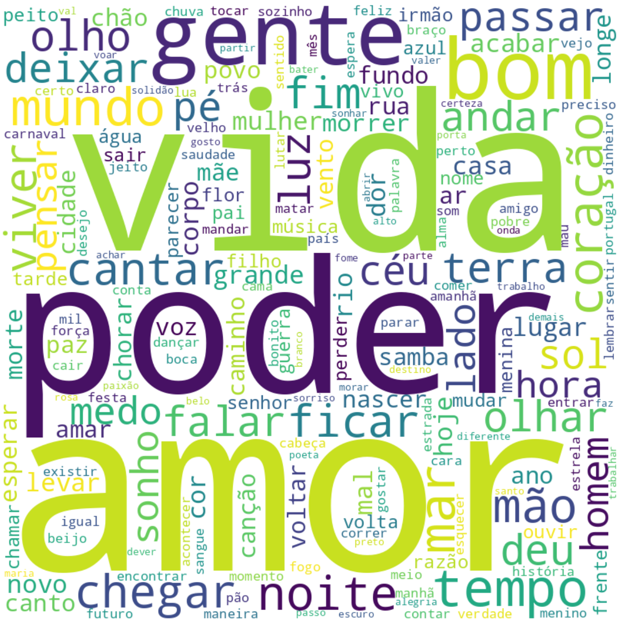


FIGURE 8. Wordcloud representation of Keywords of overall dataset

[26] https://spacy.io/usage/linguistic-features#pos-tagging (Accessed: April 2023)
[27] https://pypi.org/project/rake-nltk/ (Accessed: April 2023); Version 1.0.6

Named Entity Recognition

Named entity recognition (NER) is a fundamental natural language processing (NLP) technique often used in text mining and information extraction tasks, locating and categorizing important nouns (as organizations and locations) and proper nouns in a text (Mohit, 2014).

By using the Entity Recognizer[28] from the Spacy package, this time with the pipeline *pt_core_news_md*,[29] that presents a higher precision and accuracy on Named entity recognition tasks than the one used for the lemmatatization (Explosion, 2017) (according to the documentation, *pt_core_news_lg* has the highest evaluation values, however it did not performed as good), extracting the terms labeled as "LOC" (for location), it was strategically decided to explore the locations address by the song writers in the scope, as it was found to be interesting to have an insight if any potential streets names, cities, countries, would come up and if we would find any relations with their respective countries.

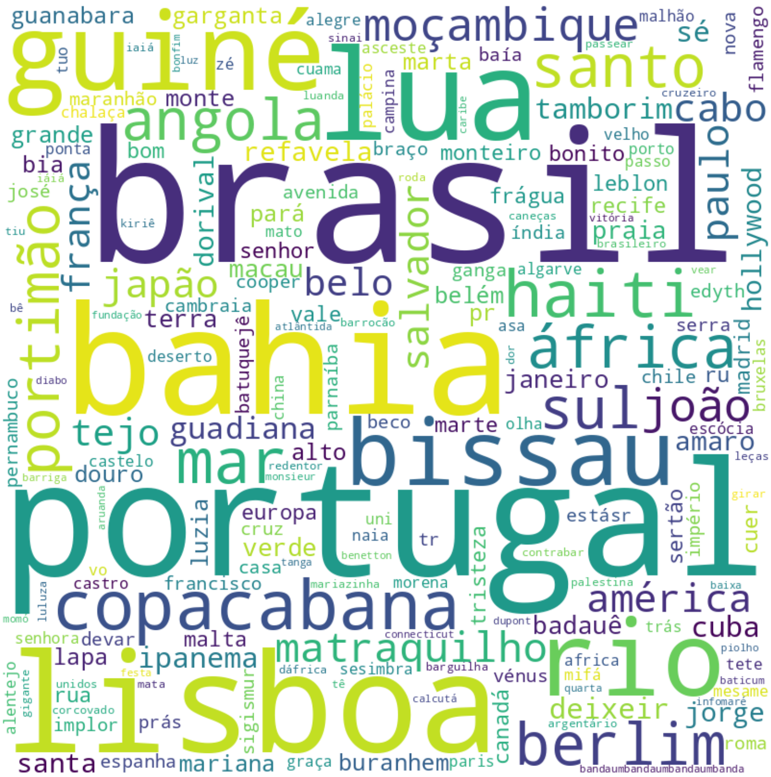


FIGURE 9. Wordcloud representation of Locations references for the entire dataset

[28] https://spacy.io/api/entityrecognizer (Accessed: April 2023)

[29] Version 3.6.0

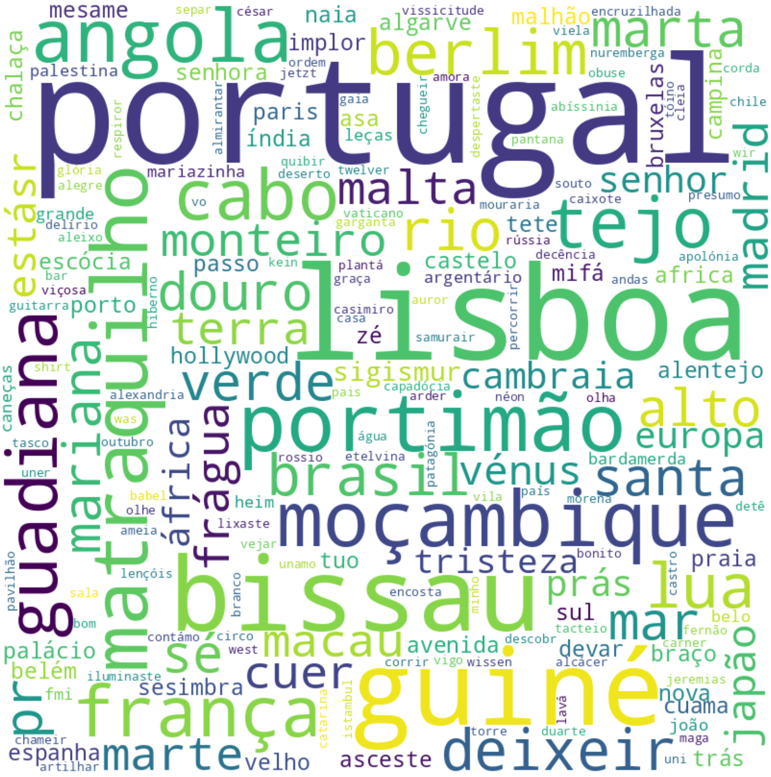


Figure 10. Wordcloud representation of Locations references found in the Portuguese song lyrics

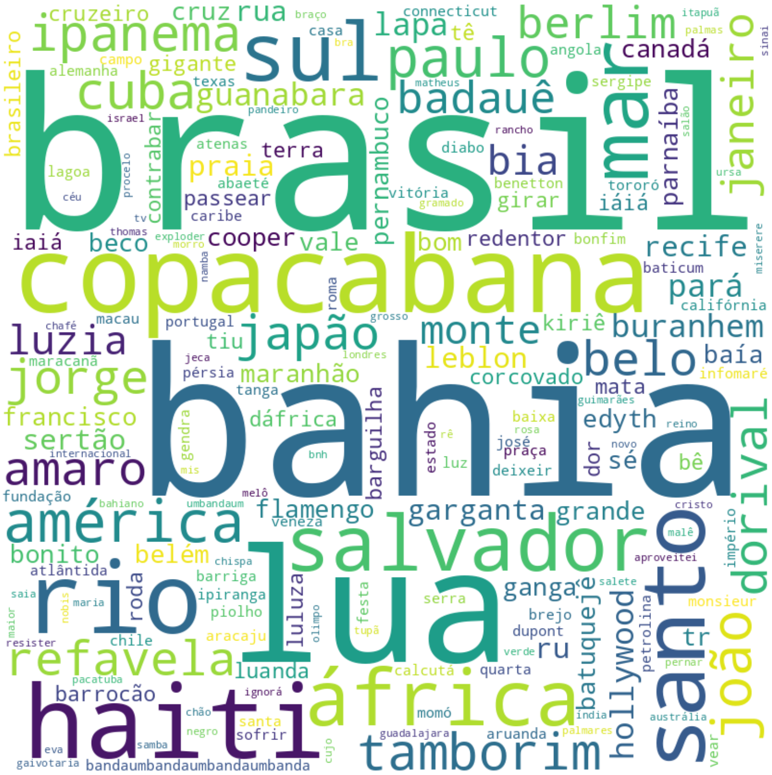


Figure 11. Wordcloud representation of Locations references found in the Brazilian song lyrics

As seen in Figures 9, 10 and 11, analysing the overall dataset without country specification, this extraction reveals a high frequency in the reference to locations such as “Bahia”, “Portugal”, “Guiné Bissau”, “Brasil” and “Rio”, with a predominance in the Portuguese lyrics of locations such as “Portugal”, “Guiné-Bissau”, “Angola”, “Portimão”, “Lisboa”, and “França”, which interrelates the nation, as it includes the capital and another Portuguese big city, and African countries that are part of the PALOP (Portuguese-speaking African countries), referring also to Portuguese rivers “Tejo”, and “Guadiana”. Similarly, the Brazilian lyrics display a high frequency of references to the Brazilian ones refer to “Bahia”, “Brasil”, “Copacabana”, “Salvador”, “Rio”, “Haiti”, and “África”, considering Brazilian cities and illustrating the closeness of Brazil and the countries in the African continent, in a historical cultural and political level (Rizzi et al., 2011).
These findings demonstrate the utility of NER in uncovering geographical and cultural references within lyrics, with distinct thematic focuses in songs from different regions.

Lexical Analysis

As part of our research, we conducted a broad lexical analysis, after the pre-processing tasks, to gain insights into how nouns, verbs, and adjectives presence in lyrics have evolved within the songs from Portugal and Brazil. To accomplish this, we employed the linguistic technique Part of Speech tagging, as explained in the preceding section. This method allowed us to precisely identify and extract words falling into the categories of nouns, verbs, and adjectives. Through this analysis, we sought to capture the dynamic linguistic trends present in the lyrical content of both countries.

The analysis showed that, although the number of Brazilian lyrics has displayed to be bigger in quantity than the Portuguese ones, the Portuguese lyrics convey an evidently larger number of words, and more diverse. For Portugal, the number of words is consistent throughout the decades, between the range of 180 and 200 words per song, while for Brazil, the word count is more volatile: in the sixties, the mean of the number of words per song rounded 131, decreasing to 80 in the seventies, manifesting an increase reaching the 126 words per song in the nineties, lowering again in the following decade, indicating a rise in the next two decades. As seen in Figures 13 and 14, the quantity of distinct words represent approximately 50% of the entire word count of the lyric for both countries.

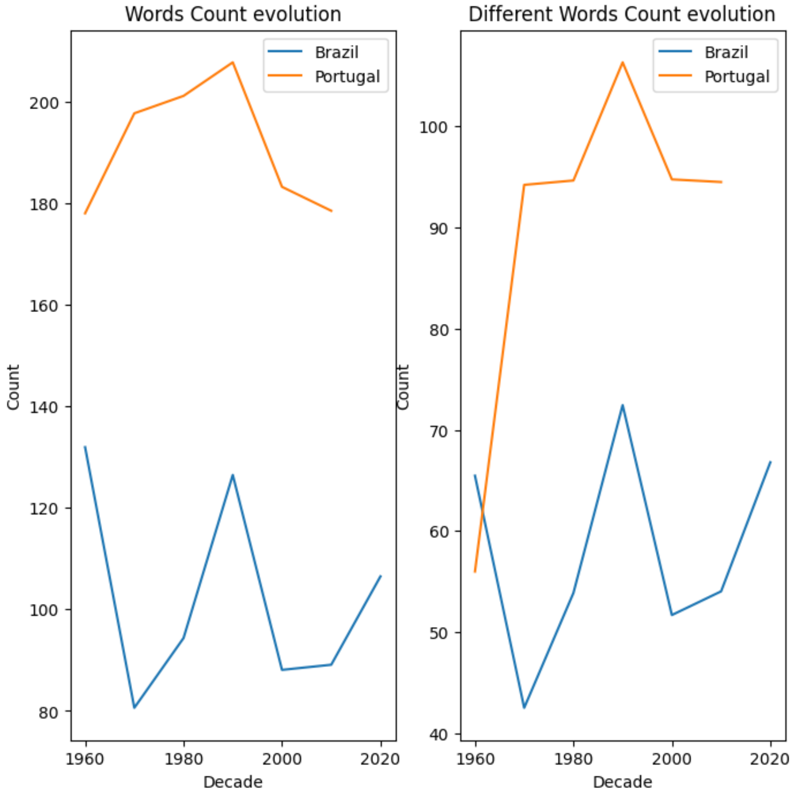


FIGURE 12. Evolution of the number of words per song

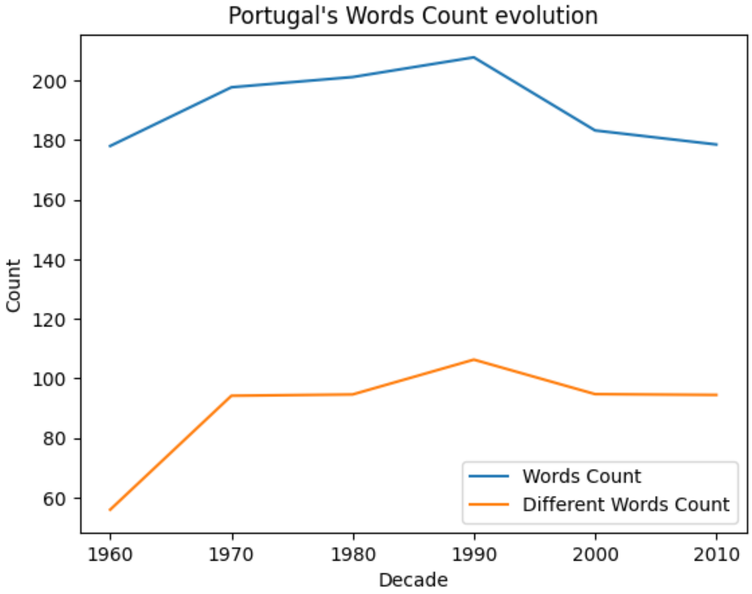


FIGURE 13. Evolution of the number of words per Portuguese song

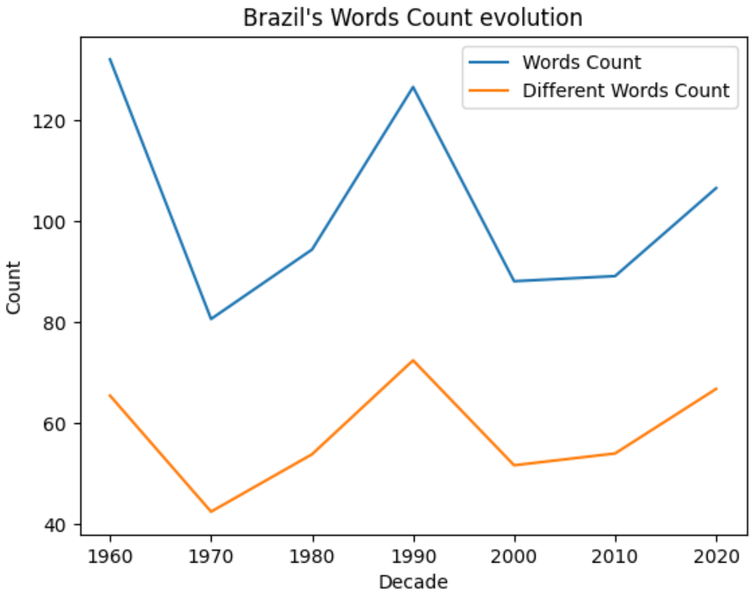


FIGURE 14. Evolution of the number of words per Brazilian song

Regarding the usage of Nouns, Verbs, Adjectives and Stop Words, the Stop Words represent the higher percentage in a song lyrics for both countries, approximately 50%, unsurprisingly, as these include all conjunctions, pronouns, which usually appear in high number. Also for both countries, the usage of Adjectives verifies to be the lowest from all four lexical categories, rounding approximately 10%. The percentages of Nouns and Verbs in the lyrics display similar behaviours, being the Nouns percentage slightly bigger, rounding 20%. Throughout the decades, the percentage of the use of these type of words have showed to be consisting.

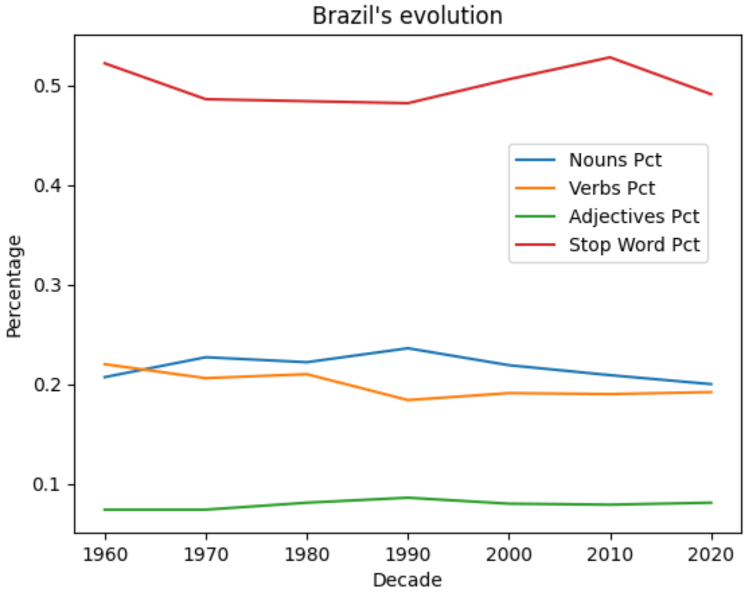


FIGURE 15. Evolution of Nouns, Verbs, Adjectives and Stop Words percentage in Brazilian song lyrics

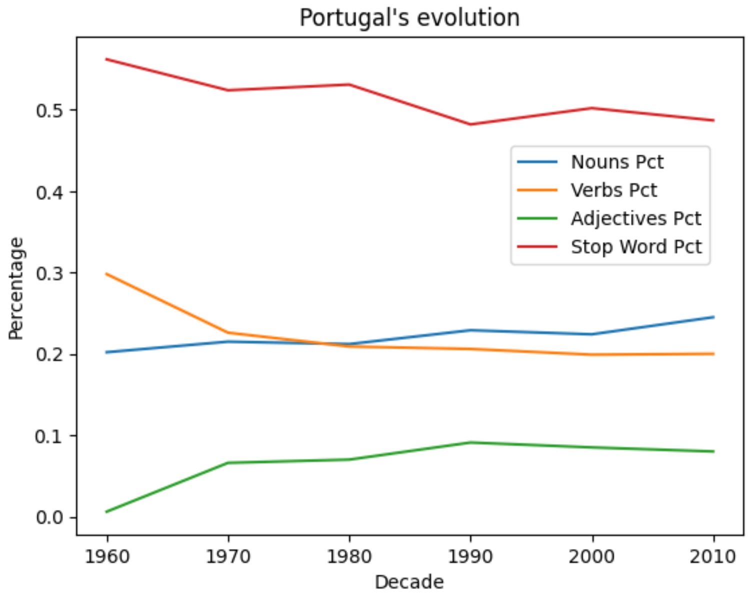


FIGURE 16. Evolution of Nouns, Verbs, Adjectives and Stop Words percentage in Portuguese song lyrics

3.4. Modeling: Topics and Sentiment Analysis

Applying the methods outlined earlier, the lyrics underwent meticulous treatment and preprocessing. This involved the creation of two distinct representations: one containing valuable bigram combinations and another that retained the text without these bigrams. Both analysis were performed on sectioned datasets, dividing between our collection on 5-year ranges, resulting in the following amount of lyrics per year range, considering valid subgroups if they included at least 10 song lyrics.[30]

Let us start with the Sentiment Analysis model.

3.4.1. Sentiment Analysis

The sentiment analysis was conducted through a lexicon-based approach, consisting of three distinct lexicons:

(1) Sentiment Polarity Lexicons[31]: this included two different corpus of positive (1399) and negative (2554) words.
(2) SentiLex-PT 02[32]: included a list of words with a sentiment polarity associated on a scale of {-1, 0, 1}, where -1 represented negative sentiment, 0 indicated

[30] Since there was only one Portuguese song in the 65's year range, this was added to the 70's, as the only 9 songs in the 2015's that were added to the 2010's group.
[31] https://www.kaggle.com/datasets/rtatman/sentiment-lexicons-for-81-languages (Accessed: July 2023)
[32] https://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3 (Accessed: July 2023)

neutrality, and 1 denoted positive sentiment. Included 1548 positive words and 4599 negative words.

(3) LIWC Dictionary (Linguistic Inquiry and Word Count)[33]: consisted in a list of 127161 terms with tags to categorize the same words. Here the terms categorized with "Negative Emotions", "Anxiety", "Anger" and "Sadness" tags were considered to have a negative sentiment and the ones with the "Positive Emotions" tag were considered to be positive. Included 848 positive words and 994 negative words.

To perform this sentiment analysis, a dictionary was meticulously constructed, incorporating the words from these lexicons and their corresponding sentiment scores. This dictionary served as the foundational reference for sentiment analysis portion of this study. For the sentiment scoring process, a custom-designed function was crafted. This function systematically traversed the pre-processed lyrics, word by word, and assigned sentiment scores based on the entries in the established dictionary. To handle negation, in the function was added a condition to flag if the word was a negative word.[34] When a negative word was found the following word's sentiment score was set to be reverted, turning a positive score into a negative and vice-versa. The words that were common between the three lexicon, were analyzed and their sentiment score were provided by the authors of this research. Through this iterative process, the lyrics were assigned to sentiment scores, enabling an analysis of emotional polarity content within the textual data contributing to a richer understanding of the nuances of positive and negative sentiment through the studied period.

Table 11 shows an example of how the scores were applied, considering the song lyrics of *Lua Cheia* by Chico Buarque.

Lyrics after pre-process for Sentiment Analysis	Sentiment Polarity Lexicon	SentiLex	LIWC	Combination of the three
chegar mar nem levar nem calar viola desconsola chorar nota ouvir voz ficar espreita espera dera abrir peito cantar feliz preparei lua cheio não não violão ficar triste pudera quisera abrir janela	[('mar', -1), ('levar', -1), ('chorar', -1), ('espreita', -1), ('feliz', 1), ('cheio', -1), ('triste', -1), ('diverso', -1)]	[(calar[34], -1), (chorar, -1), (feliz, 1), (triste, -1)]	[(viola, -1), (desconsola, -1), (chorar, -1), (espera, 1), (abrir, 1), (feliz, 1), (triste, -1), (quisera, 1), (abrir, 1)]	[(mar, -1), (levar, -1), (calar, -1), (viola, -1), (desconsola, -1), (chorar, -1), (espreita, -1), (espera, 1), (abrir, 1), (feliz, 1), (cheio, -1), (triste, -1), (quisera, 1), (abrir, 1), (diverso, -1)]
navegar mares diverso fiquei verso fiquei	Score: -6	Score: -2	Score:1	Score: -5

TABLE 11. Scores applied to song lyrics by Sentiment Polarity Lexicon, SentiLex and LIWC lexicons

[33] http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc (Accessed: July 2023)

[34] The negative words were considered "nunca", "não", "nao", "nem", "jamais".

[34] 'calar' is categorized as a positive term in this lexicon, however, being preceded by a negative word, turns its value into a negative.

3.4.2. Topic Modeling

As a next stage, we began the study and exploration of Topic Modeling, channeled into two distinct techniques: Latent Dirichlet Allocation (LDA) using the OCTIS framework and Dynamic Topic Modeling, using BERTopic.

Laying the groundwork on the OCTIS (Optimizing and Comparing Topic Models is Simple) framework, it is a tool that not only facilitates the training, analysis, and comparison of Topic Models but also revolutionizes the process by estimating optimal hyperparameters through Bayesian Optimization techniques, offering the ability to conduct fair and rigorous comparisons among their chosen topic models, leveraging benchmark datasets and renowned evaluation metrics (Terragni et al., 2021). Its code and additional information is stored in Github.[35]
For this work, OCTIS was used to canvass the best LDA model and hyperparameter optimization to obtain the topics. Our LDA modeling journey was a carefully structured process. Initializing with an exhaustive quest to identify the optimal parameter settings, crucial for the precision of our topic modeling results. Parameters such as chunk size, the number of passes and the number of topics. To gauge the effectiveness of these parameters, Coherence Score Value C_v served as the primary evaluation metric, as this is has been proven to outperform other Topic Measures (Röder et al., 2015).
Having established these foundational settings, the next effort was to pinpoint the most suitable values for *alpha* and *eta* hyper-parameters, effectively configuring our LDA model to align with the complexity of our datasets. In Table 12, there is a description of each parameter tuned and utilized.

Parameter/Hyper-parameter	Description
num_topics	The number of requested latent topics to be extracted from the training corpus
chunksize	Number of documents to be used in each training chunk
passes	Number of passes through the corpus during training
alpha	Can be set to an 1D array of lenght equal to the number of expected topics that expresses our a-priori belief for each topic's probability
eta	A-priori belief on word probability, this can be: scalar for a symmetric prior over topic/word probability, vector of length num_words to denote an asymmetric user defined probability for each word, matrix of shape (num_topics, num_words) to assign a probability for each word-topic combination, the string ‘auto’ to learn the asymmetric prior from the data

Table 12. Parameters and Hyper-paramenters descriptions' (Source: Source code for octis.models.LDA)

To lend clarity and visual depth to our topic modeling results, we used the *pyLDAvis* library, which "provides a global view of the topics (and how they differ from each other), while at the same time allowing for a deep inspection of the terms most highly associated

[35] https://github.com/MIND-Lab/OCTIS (Accessed: August 2023)

with each individual topic" (Sievert and Shirley, 2014), allowing a rich and graphical insight of the topics distributions, their top terms and their interrelationships.
After the topics were extracted from the model, a name was given to each that would describe its meaning.

Bringing an innovative approach in this research, a Dynamic Topic Modeling was also applied. This method to extract information from the topics, enabled us to dissect their evolution across our extensive dataset, spanning multiple years, revealing the adjusts and modifications of the topics over time.
Given the inherent variability and adaptability of BERTopic, multiple iterations were undertaken. These iterations aimed to identify the most fitting set of components, parameters and hyper-parameters to build and shape the model. For dimensionality reduction, UMAP (Uniform Manifold Approximation and Projection) was used, so it would be easier to visualize and cluster our documents, the lyrics. Fixating the parameter *min_ dist* as 0.0 and *metric* as cosine, we explored iterations with changes on the *n_ neighbors* and *n_ components* parameters, testing (5,10),(10,5),(10,10), respectively. Regarding the clustering parameter, we used the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) model, iterating over the *min_ cluster_ size* between 5, 8 and 10,[36] which defines the minimum number of song lyrics in each cluster. As an alternative for clustering parameter, KMeans models were also tested, with the values for *n_ clusters* of [5,10].[36] In the word universe, we used *CountVectorizer* to set the threshold of the required word frequency, setting to 3, making our model consider terms that appear in at least 5 song lyrics. To guarantee a reasonable diversity on our topics, we used Maximal Marginal Relevance, setting iterating between 0.85, 0.90, 0.95 and 1. Lastly, tuning also included the use of TF-IDF, with the *ClassTfidfTransformer* model, to mitigate the impact of words with a high frequency across the corpus, and a Sentence Transformer, using the pre-trained model *neuralmind/bert-base-portuguese-cased*.[37] Using the information and insights learned from the LDA model, we followed as human-based evaluation of the quality of the topics.

Finally, our best tuning parameters and hyper-parameters resulted in the following for Portugal:

- UMAP:
  - *n_ neighbors*=10
  - *n_ components*=5
- Clustering:
  - K-Means: *n_ components*=8
- Representation model:

[36] These values were selected based on the results from the attempts from the LDA model.
[37] https://huggingface.co/neuralmind/bert-base-portuguese-cased (Accessed: August 2023)

    - *diversity*=0.85
- Count Vectorizer
- TF-IDF
- Sentence Transformer

Being for Brazil:

- UMAP:
    - *n_ neighbors*=10
    - *n_ components*=5
- Clustering:
    - K-Means: *n_ components*=6
- Representation model:
    - *diversity*=0.85
- TF-IDF
- Sentence Transformer

3.5. Results and Evaluation: What did the Portuguese and the Brazilian feel and talked about in their songs?

In this chapter, we will present and discuss the results of our models and correlate them to the culture and history of the respective countries.

3.5.1. Sentiment Analysis

Given that different lexicons were utilized, the scores assigned to each song lyrics were differed depending on the lexicon. As seen in Figure 17 and Figure 18, the scores provided by the LIWC lexicon were all positive for Brazil, and mostly positive for Portugal, resulting in negative polarities only in the 1990's (-1). For Brazil, the remaining two lexicon do not show a large variation between them, as happens for Portugal, where the SentiLex lexicon displays values of -0.71, -2.69 and 0.17 for 1970's, 1990's and 2005's, respectively, while the Sentiment Polarity Lexicon from Kaggle, has values of -4.31, -5.77 and -1.83 for the same Year Ranges, both scoring mostly negative values. Nonetheless, all three lexicons agree that the Portuguese lyrics reached its lowest sentiment in the 1990's.

After the merge of the three lexicons, the results of the sentiments analysis appear differently. For Brazil, the songs appear to be majority positive across the studied period, except for a decrease in the 1970's and another in the 2005's years, scoring -0.71 and -0.68, respectively, showing that the songs portrait a negative sentiment, which we will further analyse, correlate and understand with the results of the topic models. Our scores indicate that the Brazilian songs reach the highest positive sentiment in the 1965's, 1980's and 2015's, being this latest the maximum with a value of 2.05. On another hand, the scores for Portugal lay more on the down side of the pole, with song conveying negative content, reaching its peak of negativity in the 1990's with a score of -5.80, which rapidly shifted in the 1995's to a score of 0.31, demonstrating a very low but yet positive score again in the 20015's. Overall, the songs from Brazil disclose a brighter and happier sentiment than the ones from Portugal, being the 2005's the only period where the the Portuguese lyrics have a higher score compared to the Brazilian ones.

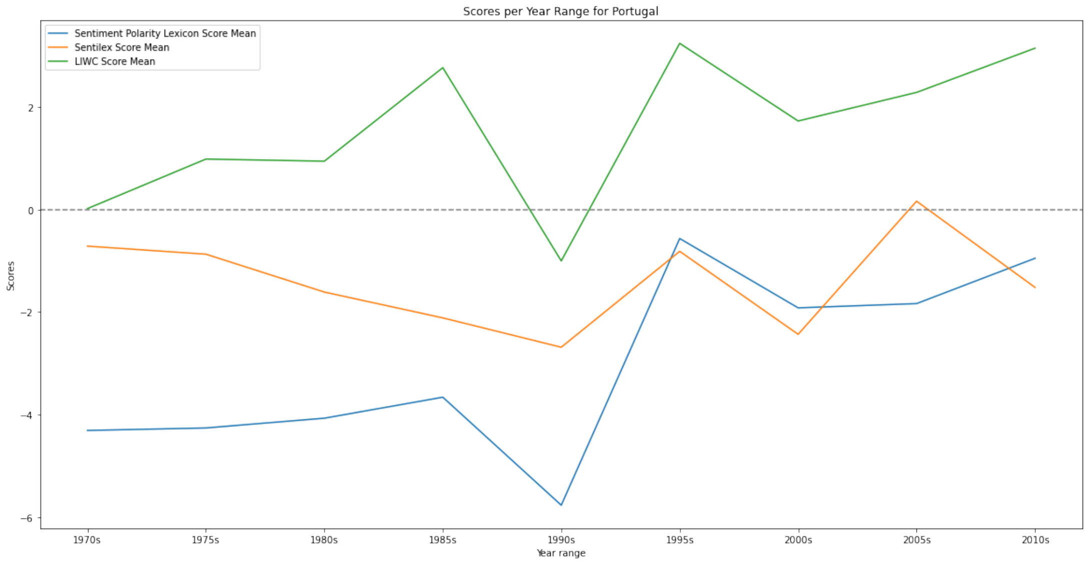


FIGURE 17. Differences in Sentiment Scores assigned to the Portuguese Song Lyrics by the 3 lexicons

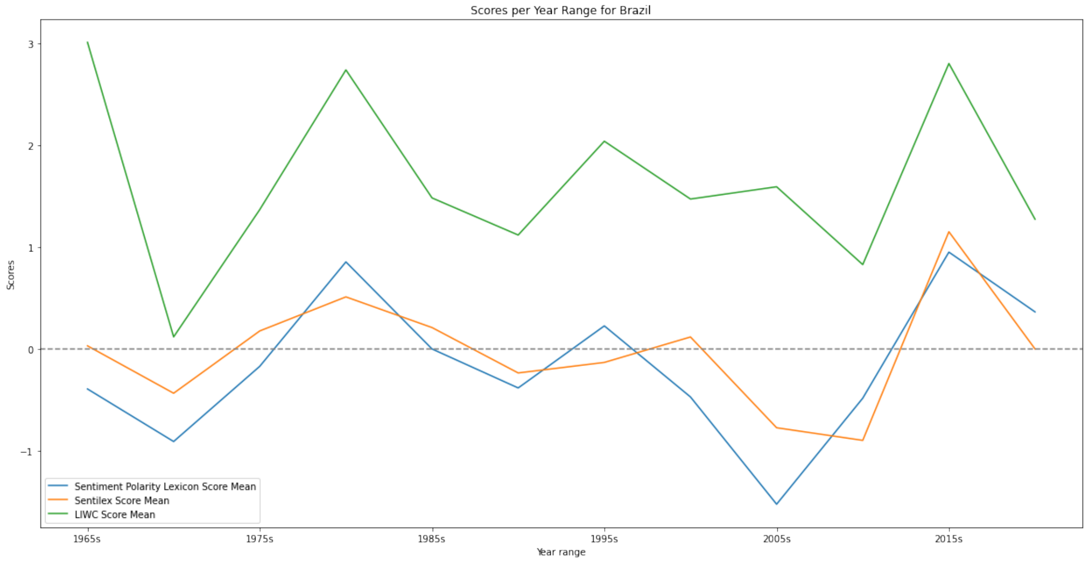


FIGURE 18. Differences in Sentiment Scores[38]assigned to the Brazilian Song Lyrics by the 3 lexicons

[38]These values were obtained by the mean of Sentiment Scores of the songs per decade.

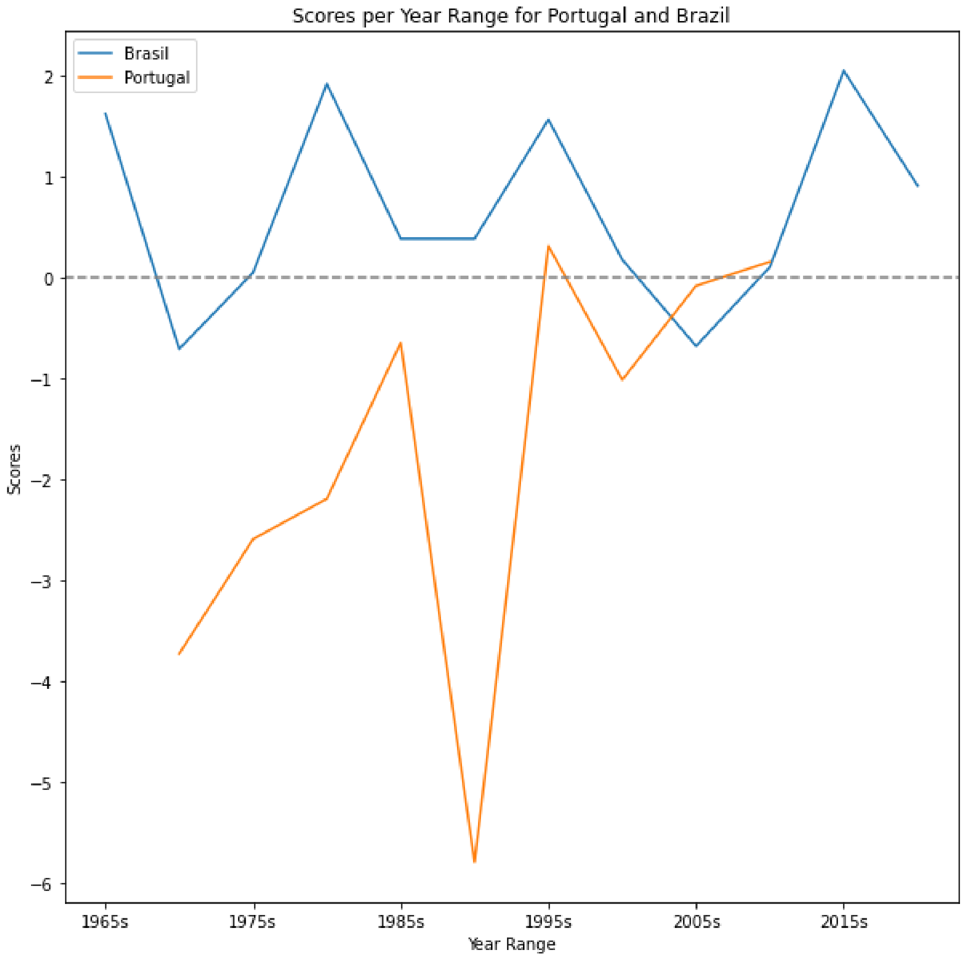


FIGURE 19. Sentiment Scores[39]assigned to the Portuguese and Brazilian Song Lyrics with the merge of all 3 lexicons

3.5.2. Topic Modeling

3.5.2.1. *Latent Dirichlet Allocation*

In this section, we will analyse the results of the model, found in Tables 13 and 14 (the "Topics Names" column includes the labels we provided to each topic, based on a semantic relationship of the topic terms provided by the algorithm), with an embedded course of the history of both country, with the goal to answer our primary question: *Is the history of the countries portrayed in the song lyrics?*

Portugal

The 1970s were a pivotal and crucial decade in the history of Portugal, marked by a series of significant and history-changing events that reshaped the country's political,

[39] These values were obtained by the mean of Sentiment Scores of the songs per decade.

social, and economic dynamics.

One of the defining events of this era was the Portuguese Colonial War, which raged on from 1961 to 1974. Also known as the War of Liberation or War of Independence, this conflict was characterized by intense armed confrontations between Portuguese armed forces and the forces organized by the African and Asian independence movements, emerging some political parties, such as People's Movement for the Liberation of Angola. The war took place primarily in Angola, Cape Verde, Guinea-Bissau, and Mozambique and played a crucial role in shaping Portugal's destiny, as it resulted in the decolonization and African independence from Portugal, for countries like Angola, Mozambique, Guinea-Bissau, Cape Verde, and São Tomé and Príncipe (Ribeiro, 2003). Through our topics results, it is visible that the songs from this period mirror this, as they refer to Freedom and African Independence topic.

On April 25, 1974, Portugal experienced a momentous turning point with a National Revolution, called the "Revolução dos Cravos", "Carnation Revolution", when translated to English, which was symbolized by the placement of carnations in the shotgun barrels of the soldiers. This non-violent military coup led to the overthrow of the long-standing authoritarian political regime, named "Estado Novo" (Ribeiro, 2003), which is highly represented in our lyrics, where we have captured the Revolution and War topics.

Economically, the 1970s posed significant challenges for Portugal. The country faced high inflation rates, mounting unemployment, and financial difficulties. In response to these adversities, the Portuguese government sought assistance from the International Monetary Fund (IMF) in 1977, receiving financial support and guidance to navigate the economic crisis. As seen in Table 13, this Fund also takes place in the topics, with the acronym FMI. The economic turbulence had a profound impact on the Portuguese society. Many citizens grappled with financial hardships, leading to concerns and anxieties about their economic well-being (Ribeiro, 2003). The 1970s were a period of resilience and adaptation for individuals and families across the nation, that we can associate with the "Austerity" and "Poverty" topics.

In the following decade, Portugal encountered revolutionary times, ambitions for better conditions and prosperous future, with the membership of the country to the European Economic Community (EEC)[40], strengthening its ties with international allies and fronts, impacting a modernization and development (Carvalho, 2017). We may see this as a reason for the motivation expressed in the songs, growing for a fight to an improvement in the justice system and better acknowledgement of the human rights, visible with the topic Democracy, justice for Human life found in this decade (Table 13).

Being the post-April 25th, in the 1980s, there was a consolidation of democracy, where regular democratic elections were held, and political stability prevailed, leading the nation to strongly affirm on the path of democratic governance (Carvalho, 2017).

[40] Eventually, integrated in the European Union, in 1993.

The 1980s were characterized by a great sense of optimism and hope for the Portuguese. After decades of political and economic instability, the population was experiencing a period of growth. Many believed that Portugal was heading in the right direction, and there was a collective belief in a brighter future (Carvalho, 2017).
However, the decade was not without its challenges. Social issues, such as unemployment and economic disparities, posed significant concerns and anxieties for the Portuguese population. These challenges underscored the importance of addressing social and economic inequalities, which is still portrait in the song lyrics of these times (Carvalho, 2017).

Regarding the 1990s and 1995s, Portugal found itself at a crossroads, motioning through notions of "grande" (big) and "pequeno" (small) on both a national and global scale, recognizing the power dynamics (Amaral, 2022). As the world continued to evolve, the Portuguese were wrestling with their own identity and place in the grand scheme of things, having accomplished large creation such as the Expo '98 (1998 Lisbon Specialised Expo), which was a specialised World's Fair held in Lisbon, themed "The Oceans, a Heritage for the Future", bringing a lot of attention to Portugal as it received over 10 million visitors (Pereira and Pereira, 2016). This environmental consciousness is also captured in our topics, as we see terms as "mar" (sea), "água" (water), "mundo" (world). Another large construction was the built of the Vasco da Gama Bridge, being 12,345 km long (Lusoponte, 2017), gaining its place as the longest bridge in the European Union.
In this period, the Portuguese also went through soul-searching, seeking for meaning and identity through its music and culture, and expressing their emotions through the same. Nostalgia and love marked this decade, with an intensity of personal relationships, being love and life experienced with fervor, which is represented in the topics. Our topics also show the sense of Passing of Time.

The decade of the 2000s in Portugal had remarks across various dimensions of Portuguese society. One of the most remarkable facets of this decade was the florescence of the culture and the performative arts, with the emergence of exceptionally talented artists, musicians, and filmmakers. These creative visionaries not only gained national recognition but also made their mark on the global stage. Portuguese culture and artistry flourished, becoming a source of inspiration for both artists and the general populace, fostering a deep appreciation for Portugal's artistic contributions. As Portuguese creativity continued to captivate audiences worldwide, the nation's cultural identity was celebrated and embraced not only within its borders but also far beyond, enhancing the "Fame" and "Entertainment" topics.
These times were leading to more positive times for Portugal. From the year of 2004, military service stopped being mandatory, and in 2005, Portugal ranked in the 19th place, as country with best life quality, in a study conducted by *Economist Intelligence Unit* (Gouveia, 2018; Lusa, 2004).

However, the global financial crisis of 2008 cast a shadow over Portugal, leading to a series of economic challenges. The country grappled with the impact of the crisis, which included a rise in unemployment rates and financial uncertainty. These economic difficulties contributed to feelings of economic anxiety and concern among the Portuguese population (Observador, 2016). Despite these challenges, Portugal remained resilient, with its cultural excellence serving as a foundation and strength during turbulent times. These conditions left a need for the Portuguese to feel that they needed to show, once again, resilience, and capacity to adapt, providing a meaning to the "Life Changing" topic.

Lastly, the 2010s marked a relief period in Portugal, characterized by diverse social, economic, and cultural shifts that deeply impacted the lives of its citizens. According to the Diário de Notícias journal (de Notícias, 2013), during this decade, many Portuguese individuals travelled abroad in search of new opportunities, symbolizing a collective desire to "escape" existing circumstances, and pursuit personal explorations to improve their and seek for new discoveries. This is shown by the "Life changing" topic, from our model, as Portugal faced significant economic and social challenges during this time, prompting fervent debates on austerity policies, employment security, and the overall state of social welfare. This brought up the sense of Security wonder and apprehensive thoughts about the economic future, as Portugal is still the country with the lowest *PIB per capita* amongst the European countries in the West (INE in Pordata, 2023).
Simultaneously, the importance of family and domestic life garnered increasing recognition throughout the 2010s. Substantial policy changes related to parental leave and a heightened focus on achieving a balance between work and family life underscored the significance of nurturing familial bonds and personal well-being (Wall et al., 2018). This emphasis on family life signified a shift in societal values, that we can also associate with the family tenderness seen in the topics (as seen in Table 13), with references to family members and cherished quality time with them.
Turning into a more liberal country, Portugal was the 6th country from Europe and the 8th in the world to legally approval marriage between a homosexual couple (Lusa, 2019). Environmental awareness surged, leading to growing concerns about "time", "rivers", "sun" and "water". This heightened environmental consciousness stemmed from a growing understanding of climate change and the imperative to safeguard the environment (Eco, 2020). Consequently, feelings of responsibility and apprehension regarding environmental issues became increasingly prevalent among the Portuguese populace. However, overall our topics demonstrate more positive times ahead.

Brazil

For Brazil, we start with the decade of 1960. A decade categorized by repression, oppression, violence and censure. At its core there were two major elements: the Military Coup of 1964 and the implementation of the repressive Institutional Act No. 5 (AI-5) in 1968 (Brückner, 2012).
The Military Coup of 1964 marked a turning point in Brazil's political landscape. It resulted in the demotion of the democratic President João Goulart, prevailing a military and dictatorship regime that would last until the year of 1985 (Brückner, 2012).
Adding on oppression, in 1968, the introduction of Institutional Act No. 5 (AI-5) brought intense rates of political violence, censure, forbidding people to speak up for their rights and beliefs, being a dark period in Brazil's history (Pitts, 2023).
These times, witnessed a cultural revolution, specially in the musical area, with movements such as "Tropicalismo" and "Jovem Guarda", counting on music lyrics to criticize the current national regime, leading to the imprisonment of multiple musicians, as was the case of Gilberto Gil and Caetano Veloso (de Macedo Duarte and de Assis César, 2021). These artists played significant roles in the struggle for freedom of expression and in exposing human rights violations during this era, having their songs served as anthems of protest (Shtromberg, 2021).
Our model demonstrate this, as we can find topics associated to military actions and restrictions, as well as music with underlying deep sad feelings.

In the following decade, the strict period persisted, with a tight control by the government, prospering an atmosphere of fear and constraint, making it relevant to highlight the terms "silêncio" and "bandeira branca" (silence and white flag, in English, respectively), as an illustration of it.
However, as the decade progressed, signs of change began to appear, with a gradual process of political opening, as the Amnesty Law at the end of the decade, as example, which allowed political exiles to return to the country and made possible the release of political prisoners, giving the Brazilian a sense of hope for a better future and to overcome the bad times, as we can see in the topic "Hoping for better times" referred for the songs in this decade (as seen in Table 14), which represented it as a transformative period (Schneider, 2021).
Even through the political issues, the country saw the appearance of a number of influential artists, musicians, and writers who gained recognition both nationally and internationally. This cultural flourishing became a source of national pride and offered an avenue for creative expression in a time of political constraint. As seen in Table 14, there is also a musical reference in the topics, with the second topic for the Year Range of the 1975s.

The topics for the decade of the 1980s reveal a happiness feeling and a renaissance of the a new nation. In fact, in the beginning of this decade, Brazil was still under the military rule, but towards the middle of the decade one of the most iconic events happened, the called "Diretas Já" (Direct Elections Now) movement, aiming to reinstate direct presidential elections and democracy, with successful results by the end of the decade, marking the official end of the military regime, laying the foundations for democracy and civil rights in the country (Schneider, 2021).
In an economic perspective, this decade was marked by a severe crisis, caused by inflation, external debt, and other economic hardships, which the government attempted to resolve with a series of economic plans (Pitts, 2023). Nevertheless, this crisis does not seem to be captured by the topics from the model: "Extraordinary musician", "Happiness" and "Brazilian regions". These seem to be associated with a positive sentiment, as we have detected in Section 3.5.1.

With the release from the military regime, the 1990s had room for political, social, and cultural developments (Pitts, 2023).
Brazil was investing in the education system, aiming to enhance its quality and accessibility of education for all, introducing important changes, as the creation of the National Education Guidelines and Bases Law in 1996 (Pitts, 2023). This importance to education was intrinsic in the song lyrics, as we can see with the "Education" topic.
However, in the decade high rates of unemployment and poverty, demonstrated with the topic referring to the "Money Loss", led to widespread discontent and protests, with Brazilians demanding more effective public policies to address these pressing issues. The struggle for social justice and equity became a central theme of the era (Sgard, 2003).

The model also showed topics of Art, Entertainment and Music, which we can associate to the diversification of Brazilian culture and media. New cultural movements emerged, contributing to a richer and more varied cultural landscape, with an international culture recognition, both in the musical and film industries.

Entering the twenty first century, Brazil was governed by a new president, bringing a boom in the economic area, a growth, with a relief from poverty and an investment (Pitts, 2023). However, our results do not reflect that. The topics seem to be related to Sea and Sea traveling, which we can potentially relate to the significant oil reserves discoveries in the pre-salt layer off the Brazilian coast. This discovery contributed to the increase of the economic potential for the nation.

At the break of the 2010s, Brazil officially became a part of the BRIC, along with other international great powers, Russia, India and China, empowering its economic presence in across nations, correlating with the Empire-related topic.

The other topics in this decade convey a cherishing and tender message, which actually contrast with the Brazilian experience at the end of this period. With the election of Jair Bolsonaro as president, in 2018, the country shifted to a conservative environment, both on social and economic matters.

The results we get from the 2020s are not very representative, since we only counted with 11 songs in this Year Range, being 10 of them from Caetano Veloso. Our topics indicate the Music, referring to renowned figures like Jorge Ben, Pixinguinha, and Djavan who have continued to inspire and delight audiences, both within Brazil and globally, transcends generations, and embracing good life moments, despite the uncertainties and difficulties faced in the this specific period (Reily, 2000).
When facing adversities, Brazilians have found consolation in their artistic and cultural heritage, embracing life's simple pleasures and cherishing the beauty of their land, demonstrating the resilience that has defined their nation's spirit for generations.

Year Range	Nr topics	c_v score	Topics terms	Topics Names
1970's	4	0.591	[[olho, noite, vida, gente, mariazinha, tornei, marta, pràquilo, verde, mão],	topic 1: undefined[41]
			[chegar, homem, vento, tocar, cantiga, cantar, gente, segar, rosa, partir],	topic 2: People coming and going
			[senhor, bom, matar, soldadinho, trabalhar, terra, menino, guerra, burguês, aprender],	topic 3: Fighting and war
			[vida, ficar, mundo, marcolino, mudar, parar, arma, noite, perder, medo]]	topic 4: Revolution
1975's	4	0.575	[[quadra, amigo, vejo, pobre, senhor, homem_fantasma, fraquinho, dás, contar, vento],	topic 1: undefined[41]
			[guiné_bissau, livre, independente, angola, moçambique, senhora, vitória, trabalhador, viva, cantar],	topic 2: African countries and freedom
			[guarda, velho, preço, pé, comprar, saia, aguentar, fome, pode, maria],	topic 3: Poverty
			[filho, fmi, gente, vida, mão, poder, casa, culpa, amor, terra]]	topic 4: Austerity times
1980's	2	0.547	[[futuro, duro, sol, olhar, sonho, lembrei, mostrar, luz, ouro, fogo],	topic 1: Hoping for future times
			[vida, abraço, cola, linda, tempo, mil, partida, pedaço, gente, terra]]	topic 2: undefined[41]
1985's	2	0.635	[[mundo, direito_humano, mãe, democracia, mau, sistema, abel, exceção, caim, país],	topic 1: Democracy, justice for Human life
			[amor, chover, dor, tempo, noite, perder, vida, parecer, chuva, olhar]]	topic 2: Pain and Loss
1990's	5	0.686	[[grande, poder, pequeno, tamanho, pequenino, mão, pedra, noz, laje, asfalto],	topic 1: Dimention
			[sentido, rua, canção, único, saudade, cantar, nenhum, porta, circulação, ligar],	topic 2: Musical emotions
			[tempo, passar, sol, cem, novo, vivo, noite, folha, fico, mil],	topic 3: Passage of time
			[amor, dairinhas, vida, amar, paixão, pai, azul, conhecer, homem, lei],	topic 4: Love
			[mar, água, cair, céu, lançar, coração, morte, mundo, vento, andar]]	topic 5: Environmental Awareness
1995's	2	0.630	[[amor, frequente, vida, arma, ardente, respirar, olhar, menina, pé, favor],	topic 1: Love and Life
			[capote, branco, preto, cor, defender, xico, convém, usa, flanco, lado]]	topic 2: Racial dispute
2000's	4	0.405	[[tempo, bom, lisboa, cantar, sair, amor, olho, bairro_alto, contar, novo],	topic 1: Lisbon entertainment and social
			[pão, amor, prisão, cantor, morrer, vida, flor, cantar, sozinho, demitir],	topic 2: Loneliness
			[grande, poder, anão, portugal, fundo, chegar, mundo, pé, tama, história],	topic 3: Portugal's history
			[lixo, casa, azul, gente, sjoão, senhor, rei, achar, poder, praxe]]	topic 4: Overpowering
2005's	5	0.497	[[tempo, rosa, vermelho, alerta, luz, poder, chegar, viva, tomar_conta, feliz],	topic 1: undefined[41]
			[velho, samurair, cinza, luz, poder, amar, olha, levar, vencer, tempo],	topic 2: Gracefully winning fight; Power
			[país, famoso, rei, eleger, grande, fama, mal, verdade, nome, perfeito],	topic 3: Fame and hierarchy
			[circo, monteiro, pé, marcha, luz, lona, tempo, caixa, amar, novo],	topic 4: Entertainment
			[vida, valor, viver, mudar, jmb, alc, mão, fabricar, roda, razão]]	topic 5: Life changing
2010's	5	0.712	[[guarda, filhinha, pano, armário, cru, aguarela, baú, caixa, exato, gargalhada],	topic 1: Entertaining activities with loved ones
			[vida, amor, fugir, tocar, sangue, jogo, parte, faz_parte, audácia, retorno],	topic 2: Life changing, coming back
			[filhinha, guarda, pano, armário, mãezinha, gargalhada, exato, lata, cru, caixa],	topic 3: same as topic 1
			[acesso, bloquear, adivinhar, presente, dever, mal, ficar, seguro, lado, passado],	topic 4: Secret
			[música, medida, cabe, tamar, tempo, rio, rodar, roda, sol, água]]	topic 5: Nature

TABLE 13. Topics extracted from the LDA topic model for Portuguese Song Lyrics

[41] There is no correlation between the terms, making it challenging to label the topic.

Year Range	Nr topics	c_v score	Topics terms	Topics Names
1965's	2	0.430	[[esperar, copacabana, proibir, sol, ano, viva, clarin, banda_militar, irene, sozinho],	topic 1: Hoping for end of military actions
			[amor, samba, mundo, cantar, ficar, mar, coração, noite, chorar, passar]]	topic 2: Music as way of expression
1970's	3	0.653	[[praça, frevo, chegar, gente, novo, tempo, mundo, céu, pintar, perder],	topic 1: Hoping for better times
			[bandeira_branco, silêncio, pé, saia, arriba, enfiar, pau, triste, forte, mar],	topic 2: Subdued
			[gil, gendra, rouxinol, mulher, mulato, balangandãs, passar, buranhem, vadeia, vadiar]]	topic 3: Lifestyle
1975's	1	0.509	[[cabeça, asa, mau, avesso, novo, chamar, gosto, bom, frente, sonho],	topic 1: Hoping for better times
			[tempo, terra, gente, amor, cantar, bom, menina, dançar, coração, esquecer]]	topic 2: Music, dancing
1980's	2	0.496	[[linda, amor, canção, homem, demais, estrela, deixa, peter_gast, gosto, forte],	topic 1: Extraordinary musician
			[mar, amor, luz, azul, feliz, vida, onda, música, quer, beleza]]	topic 2: Happiness
1985's	2	0.677	[[comer, falar, baiano, pára, trinidad, joão, poesia, nascer, crescer, ouvir],	topic 1: Brazilian regions
			[gente, amor, bom, poder, noite, luz, tempo, azul, feliz, branco]]	topic 2: undefined[42]
1990's	5	0.681	[[samba, cinema, novo, filme, ordem, grande, conversa, palavra, luz, nome],	topic 1: Art and Enternainment
			[dorival, mundo, linchador, camar, tempo, volta, cujo, nação, triste, fruto],	topic 2: Reflections on National Identity
			[americano, preto, branco, pobre, perder, mulher, fundo, dinheiro, algo, quebrar],	topic 3: Money loss
			[lua, televisão, certo, santa, possa, clara, deus, calar, nu, sinal],	topic 4: Spirituality, god belief
			[brau, preta, nobreza, cidade, salvador, liberdade, elegante, palavra, sinal, trabalho]]	topic 5: City life
1995's	3	0.473	[[céu, menino, alexandre, pai, felipe, aprender, hora, ciência, nascer, grande],	topic 1: Education
			[povo, pai, vento, tempo, tak, lugar, mar, cor, eta, chão],	topic 2: undefined[42]
			[cantar, tom, manhattan, deixar, carnaval, viver, poder, menina, coração, mar]]	topic 3: Music and party
2000's	2	0.971	[[reza, poder, amor, levar, vela, certo, rima, maré, mirar, rema],	topic 1: Sea travelling
			[grande, bahia, rio, hoje, axé, curuzu, dodô, maconha, éter, wyoming]]	topic 2: undefined[42]
2005's	3	0.727	[[mar, medo, poder, amor, abrir, parar, intenso, cal, deuses, imenso],	topic 1: Fear of sea travelling
			[tarar, tarado, grude, mundo, carnaval, nu, gostar, deixa, marca, lua],	topic 2: Carnal pleasure
			[lua, olho, lapa, branco, talho, malho, luz, brilho, chuva, pele]]	topic 3: undefined[42]
2010's	3	0.602	[[comunista, baiano, mulato, terra, horror, morrer, homem, sonho, guardar, tédio],	topic 1: Conflict
			[mar, mundo, chegar, coração, amor, olho, exasperar, raiva, império, lei],	topic 2: Ruthless empire
			[vida, viver, amor, tempo, joão, reina, cantar, noite, cor, sol]]	topic 3: Enjoying life
2015's	2	0.555	[[coração, vovó, mamãe, ouço, pensar, novo, yamandu, poder, pedacinho, mandar],	topic 1: Tenderness
			[amor, tamanho, bom, coração, poder, lia, deia, noite, axé, pé]]	topic 2: Good intentions
2020's	5	0.782	[[tom, querubim, deslumbramento, autoacalanto, benjamim, ninar, curumim, lamento, emular, ja],	topic 1: Music
			[jorge, ben, pixinguinha, ensinar, português, som, sentido, djavans, pousa, canto],	topic 2: Music Language
			[tela, tempo, vale, azul, anjo, império, silício, mi, poder, troncho],	topic 3: undefined[42]
			[deixar, amor, chegar, cantar, samba, gente, vovô, luta, gozo, esculachar],	topic 4: Enjoying Life
			[vale, viver, bronze, nobre, tez, mulher, céu, sol, amor, santana]]	topic 5: Nature Views

Table 14. Topics extracted from the LDA topic model for Brazilian Song Lyrics

[42] There is no correlation between the terms, making it challenging to label the topic.

3.5.2.2. *Dynamic Topic Modeling*

As stated in Section 2.2.2, the DTM uses BERTopic as a foundation, from which, we have extracted the topics by the top 4 terms representative of each, which we then labeled as found in Table 15.

Country	Topic key	Topics terms	Topics Names
Portugal	0	[coração, andar, hino, mão]	Patriotism
Portugal	1	[luz, mar, ar, doce]	Nature and Serenity
Portugal	2	[filho, tempo, contar, poder]	History of Power
Portugal	3	[caminho, mão, irmão, lugar]	Journey
Portugal	4	[frágil, guerra, acesso, homem]	Nation fragility
Portugal	5	[portugal, capital, gente, poder]	Portuguese system
Portugal	6	[bairro, fugir, tocar, emprego]	Seek for Life Meaning and Better opportunities
Portugal	7	[música, tocar, malta, senhor]	Music
Brazil	0	[gil, rouxinol, xangô, silêncio]	Spirituality
Brazil	1	[tempo, terra, esperar, azul]	Hope
Brazil	2	[amor, medo, grude, poder]	Fear and Power
Brazil	3	[ficar, logo, volta, noite]	People arrivals and departures
Brazil	4	[música, poeta, rádio, peter_gast]	Music
Brazil	5	[sambar, traga, esquece, deixe]	Music expressing memories

Table 15. Topics extracted from the BERTopic model

Starting with Portugal, we can observe (in Figure 20) that the topic 0 seems consistent throughout the years, with a spike in early 1980s. Nature seems to have a rise in the 2010s, confirming our results in with the LDA model. Regarding the Power, we are presented with clear higher results up to the year of 1985, which can be associated with references to the war and all the extreme power that was conducted until 1975, with a similar behaviour as the Topic 5, which is referring to the Portuguese system and its dynamics. The Music topic, shows a spike near the 2005s, which we can correlate with the previously analyzed Entertainment and Social topic, for this period with the LDA model. Lastly, Topic 6 shows higher frequencies up until the 1990s, however, the frequencies do not get very low during the period, demonstrating the need of the Portuguese to seek else where for ways to improve their lives, for example for better job opportunities.

For Brazil, as seen in Figure 21, spirituality seems to carry a lot of weight, starting to decrease in the break of the 21st century, similarly as the Hope topic, which can be associated with the times as when the Poverty started to decrease in the country. For the Fear and Power topic, its maximal value is in the sixties, which seems to mirror the era of the Military Coup, as seen before. Contrasting with the LDA topics, Music here seems to be lowly represented, with high values only in the 1975s. Finally, the last topic, has its appearance only in the 1965s, which was also seen with the LDA model.

In the Appendix chapter the graphic distribution per topic can be found.

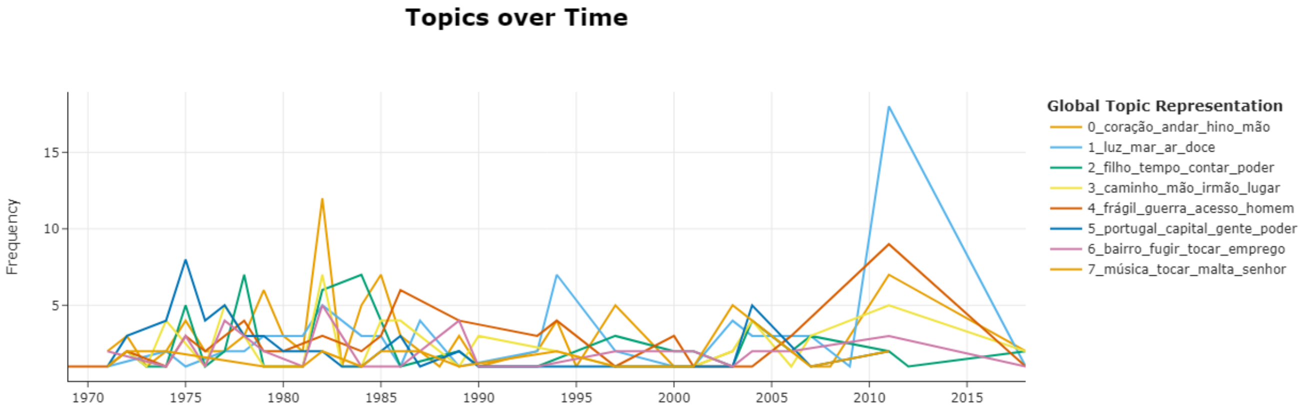


FIGURE 20. Dynamic Topic Distribution for Portuguese Song Lyrics

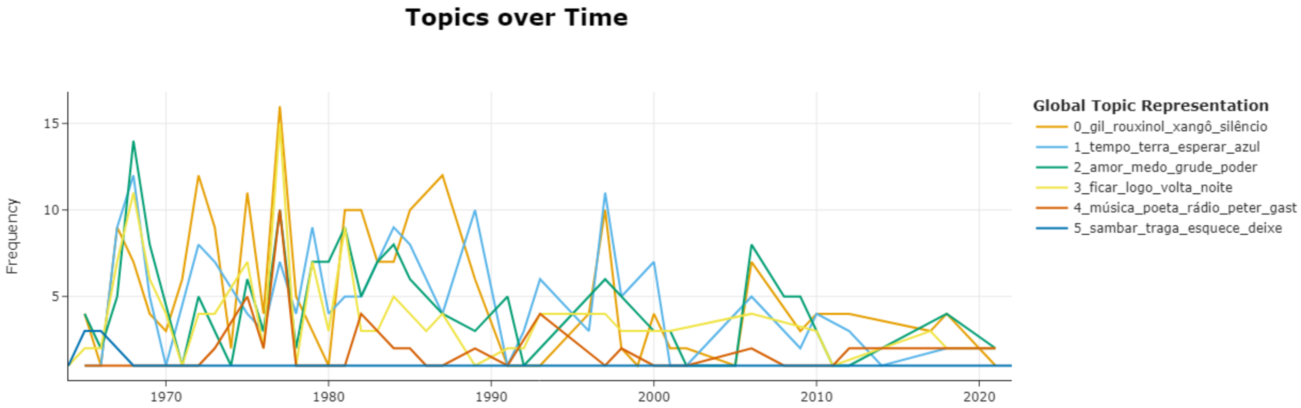


FIGURE 21. Dynamic Topic Distribution for Brazilian Song Lyrics

Overall, our results, demonstrate that even through rough phases in their nation history, the Brazilian song lyrics appear to convey a more positive message and sense of hope.

CHAPTER 4

Conclusion

In conclusion, this study has reached some interesting patterns and results. Firstly, it was notable the contrast in the emotional tone between the two countries' music, as Brazilian songs tend to lean towards a more positive and uplifting sentiment, while Portuguese songs often carry a predominant negative emotional undertone.

We found that our exploration of the extracted topics from these lyrics generally aligns with the historical and societal experiences of each nation, however, there are instances where the lyrics do not accurately reflect the challenging periods, neither on the topics or the emotion polarities, which hints that these lyricists may also use their musical poems as an escape to the reality.

Referring back to our pre-analysis researches, it is also worth noting that our findings challenge the conventional notion that negativity is more prevalent in recent times. Instead, we observe a dynamic and fluctuating pattern in Brazilian music lyrics over the years, with periods of both positivity and negativity. In the case of Portugal, we actually witness a trend towards a more positive emotional tone in recent years.

This research highlights the rich dynamics of emotions and experiences encapsulated in the music of Brazil and Portugal.

To complete this research, we have encounter multiple limitations and challenges. Starting with the process of discovering existing studies focusing on the Portuguese language, this revealed to be a true challenge, as most of the palette of researches done in this field for this goal, are for English and Indian idioms, which turned into another challenge when it came to finding the best models that would suit the Portuguese language, requiring pioneering efforts in data analysis.

The acquisition of the Portuguese song lyrics, also required an intensive process of manually scouting through album booklets and scanning extensive books with numerous pages, as there was not an existing official digital source of these texts.

Still related to the Portuguese language, is how well the text mining models are prepared for it. We have faced certain limitations on this, for example is the Part of Speech tagging, resulting in compromises in the analytical phase, facing a non-great accuracy. The sentiment analysis could have also been compromised, due to the lack of diversity of words within the lexicon.

When dealing with the quality of the topics, the subjectivity played its role, turning the process more intriguing, as it is important to consider a certain level of sensitive in its

interpretation.
Lastly, from the technical and computational perspective, we came across its own set of limitations. Issues related to package compatibility and version management often required resourceful problem-solving, while the constraints regarding computational resources in the Google Colab tool also presented recurrent set backs, necessitating creative strategies to address them.

These limitations served to enrich the complexity of the experience, however, they were an opportunity to improve a meticulous planning process as well, expanding the resources and adaptability.

In the context of future work for this area, there are several promising paths to explore that can enhance and improve such analysis. In the Data Preparation phase, using part-of-speech tagging to combine a more robust list of stop words, extracting words of auxiliar, coordinating conjunction, determiner and interjection classes. An investment in more and different Sentence Transformers, could offer refining tuning for the topic models, presenting interesting results. Secondly, the incorporation of the Multilanguage model to further analyze the emotional content of lyrics in a multilingual context.
Additionally, for a more holistic analysis, the inclusion of the lyricist age as a variable can add a layer for analysis. By considering the age of lyricists in relation to the lyrics they produce, we can gain insights into how the age of the writer shapes their lyrics, expanding even the lyricist selection to capture multiple with different ages, writing about the same theme.

References

Agrawal, Y., Shanker, R. G. R., & Alluri, V. (2021). Transformer-based approach towards music emotion recognition from lyrics. In D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, & F. Sebastiani (Eds.), *Advances in information retrieval - 43rd european conference on IR research, ECIR 2021, virtual event, march 28 - april 1, 2021, proceedings, part II* (pp. 167–175). Springer. https://doi.org/10.1007/978-3-030-72240-1_12

Ahuja, M., & Sangal, A. (2018). Opinion mining and classification of music lyrics using supervised learning algorithms. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 223–227.

Amaral, L. (2022). *Economia portuguesa: Últimas décadas*. Fundação Francisco Manuel dos Santos.

Amin, M., Fankhauser, P., Kupietz, M., & Schneider, R. (2021). Data-driven identification of idioms in song lyrics. *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, 13–22.

Azevedo, A., & Santos, M. F. (2008). Kdd, semma and crisp-dm: A parallel overview. *IADS-DM*.

Aziz, R. A., Bijaksana, M. A., et al. (2019). Two stages song subject classification on indonesian song based on lyrics, genre & artist. *2019 7th International Conference on Information and Communication Technology (ICoICT)*, 1–6.

Baas, J., Schotten, M., Plume, A. M., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quant. Sci. Stud.*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Bai, X., Zhang, X., Li, K. X., Zhou, Y., & Yuen, K. F. (2021). Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*, *102*, 11–24.

Barman, M. P., Awekar, A., & Kothari, S. (2019). Decoding the style and bias of song lyrics. In B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, & F. Scholer (Eds.), *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2019, paris, france, july 21-25, 2019* (pp. 1165–1168). ACM. https://doi.org/10.1145/3331184.3331363

Besson, M., Faita, F., Peretz, I., Bonnel, A.-M., & Requin, J. (1998). Singing in the brain: Independence of lyrics and tunes. *Psychological Science*, *9*(6), 494–498.

Betti, L., Abrate, C., & Kaltenbrunner, A. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, *12*(1), 10.

Blei, D., Ng, A., & Jordan, M. (2001). Latent dirichlet allocation. *Advances in neural information processing systems*, *14*.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, 113–120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bradlow, E. T., & Fader, P. S. (2001). A bayesian lifetime model for the "hot 100" billboard songs. *Journal of the American Statistical Association*, *96*(454), 368–381.

Brand, C. O., Acerbi, A., & Mesoudi, A. (2019). Cultural evolution of emotional expression in 50 years of song lyrics. *Evolutionary Human Sciences*, *1*.

Brückner, J. (2012). From military dictatorship to democratic consolidation.

Carvalho, M. (2017). *Quando portugal ardeu*. Leya.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, *9*(13), 1–73.

Choi, J., Song, J., & Kim, Y. (2018). An analysis of music lyrics by measuring the distance of emotion and sentiment. *19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2018, Busan, Korea (South), June 27-29, 2018*, 176–181. https://doi.org/10.1109/SNPD.2018.8441085

Choi, K., Lee, J. H., Willis, C., & Downie, J. S. (2015). Topic modeling users' interpretations of songs to inform subject access in music digital libraries. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 183–186.

Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, *54*(10s), 1–35.

Cole, R. R. (1971). Top songs in the sixties: A content analysis of popular lyrics. *American Behavioral Scientist*, *14*(3), 389–400.

Costa, A. R. d. A. d. (2017). *A text-mining based model to detect unethical biases in online reviews: A case-study of amazon.com* (Master's thesis).

Dang, T.-T., & Shirai, K. (2009). Machine learning approaches for mood classification of songs toward music search engine. *2009 International Conference on Knowledge and Systems Engineering*, 144–149.

de Lucio, J., & Palomeque, M. (2022). Music preferences as an instrument of emotional self-regulation along the business cycle. *Journal of Cultural Economics*, 1–24.

de Macedo Duarte, A., & de Assis César, M. R. (2021). Corpos, gêneros e sexualidades em disputa no brasil contemporâneo: Bolsonarismo versus tropicalismo. *História: Questões & Debates*, *69*(2), 75–95.

de Notícias, D. (2013). https://www.dn.pt/portugal/emigracao-cresceu-85-entre-2010-e-2011-%202999213.html

Devika, M., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science*, *87*, 44–49.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DeWall, C. N., Pond Jr, R. S., Campbell, W. K., & Twenge, J. M. (2011). Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular us song lyrics. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(3), 200.

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics*, *125*, 37–46.

Eco. (2020). 76% dos portugueses estão preocupados com o ambiente. https://eco.sapo.pt/2020/06/05/76-dos-portugueses-estao-preocupados-com-o-%20ambiente-e-alteracoes-climaticas/

Explosion. (2017). Portuguese · spacy models documentation. https://spacy.io/models/pt

Frith, S., Straw, W., & Street, J. (2001). *The cambridge companion to pop and rock*. Cambridge University Press.

Godinho, S., & Saraiva, A. (1983). *Cançoes de sérgio godinho*. Assirio e Alvim.

Gouveia, J. B. (2018). *Direito da segurança*.

Grootendorst, M. P. (2021). Dynamic topic modeling - bertopic. *maartengr.github.io*. Retrieved September 2023, from https://maartengr.github.io/BERTopic/getting_started/topicsovertime/topicsovertime.html#visualization

Gupta, T. (2017). Keyword extraction: A review. *International Journal of Engineering Applied Sciences and Technology*, *2*(4), 215–220.

Habert, B., Adda, G., Adda-Decker, M., de Mareüil, P. B., Ferrari, S., Ferret, O., Illouz, G., & Paraubeck, P. (1998). Towards tokenization evaluation. *Lrec*, 427–432.

Hearst, M. (2003). What is text mining. *SIMS, UC Berkeley*, *5*.

Jo, W., & Kim, M. J. (2023). Tracking emotions from song lyrics: Analyzing 30 years of k-pop hits. *Emotion*, *23*(6), 1658.

Kamalnathan, S., Mishra, Y., Kumawat, V., & Bangwal, V. (2019). Evolution of different music genres.

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Kharis, M., Kisyani, Suhartono, Pairin, U., & Darni. (2021). How to lemmatize german words with nlp-spacy lemmatizer? *Proceedings of the International Seminar on Language, Education, and Culture (ISoLEC 2021)*, 189–193. https://doi.org/10.2991/assehr.k.211212.036

Kherwa, P., & Bansal, P. (2020). Topic modeling: A comprehensive review. *EAI Endorsed Trans. Scalable Inf. Syst.*, *7*(24), e2. https://doi.org/10.4108/eai.13-7-2018.159623

Kim, M., & Kwon, H.-C. (2011). Lyrics-based emotion classification using feature selection by partial syntactic analysis. *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, 960–964.

Laoh, E., Surjandari, I., & Febirautami, L. R. (2018). Indonesians' song lyrics topic modelling using latent dirichlet allocation. *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, 270–274.

Laokok, S., & Khonthapagdee, S. (2022). Emotion classification in thai music using convolutional neural networks. *2022 6th International Conference on Information Technology (InCIT)*, 148–151.

Lee, C.-S., Wang, M.-H., Chen, L.-C., Lai, S.-Y., & Kubota, N. (2018). Fuzzy semantic agent based on ontology model for chinese lyrics classification. *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 4254–4259.

Lusa. (2004). "the economist": Portugal é o 19 o país com melhor qualidade de vida. https://www.publico.pt/2004/11/18/sociedade/noticia/the-economist-%20portugal-e-o-19%C2%BA-pais-com-melhor-qualidade-de-vida-1208724

Lusa. (2019). Faz hoje 9 anos que entrou em vigor em portugal a lei do casamento entre pessoas do mesmo sexo. https://comunidadeculturaearte.com/faz-hoje-9-anos-que-entrou-em-vigor-em-%20portugal-a-lei-do-casamento-entre-pessoas-do-mesmo-sexo/

Lusoponte. (2017). https://www.lusoponte.pt/vasco-da-gama/projecto/principais-caracteristicas

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015). Sentiment analysis techniques in recent works. *2015 science and information conference (SAI)*, 288–291.

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016). Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, *9*(2), 240–254.

Martinez, A. R. (2012). Part-of-speech tagging. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(1), 107–113.

Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific and Technology Research*, *9*(2), 601–609.

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department.*

Mohit, B. (2014). Named entity recognition. In *Natural language processing of semitic languages* (pp. 221–245). Springer.

Napier, K., & Shamir, L. (2018). Quantitative sentiment analysis of lyrics in popular music. *Journal of Popular Music Studies*, *30*(4), 161–176.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653–7670.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd international conference on Knowledge capture*, 70–77.

Nath, D., & Phani, S. (2021). Mood analysis of bengali songs using deep neural networks. In *Information and communication technology for competitive strategies (ictcs 2020)* (pp. 1103–1113). Springer.

Nokel, M., & Loukachevitch, N. (2015). Topic models: Accounting component structure of bigrams. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, 145–152.

Observador. (2016). Crise e castigo. a longa estagnação da economia em portugal. https://observador.pt/especiais/crise-castigo-longa-estagnacao-da-economia-%20portugal/

Oudenne, A. M., Swarthmore, P., & Chasins, S. E. (2010). Identifying the emotional polarity of song lyrics through natural language processing. *Swarthmore, EUA, sd 13p. Relatório final apresentado no CPSC*, *65*.

Palma, J., & Callixto, J. C. (2005). *Na terra dos sonhos*. Quasi.

Pereira, A. M., & Pereira, R. M. (2016). Investimentos em infraestruturas em portugal. *Lisbon, Portugal: Fundação Francisco Manuel dos Santos*.

Perera, P., & Witte, R. (2005). A self-learning context-aware lemmatizer for german. *Human Language Technology - The Baltic Perspectiv*.

Pettijohn, T. F., & Sacco Jr, D. F. (2009). The language of lyrics: An analysis of popular billboard songs across conditions of social and economic threat. *Journal of language and social psychology*, *28*(3), 297–311.

Pitts, B. (2023). *Until the storm passes*. Univ of California Press.

Pordata. (2023). https://www.pordata.pt/portugal/taxa+de+crescimento+do+pib-2298

Rafi-Ur-Rashid, M., Mahbub, M., & Adnan, M. A. (2022). Breaking the curse of class imbalance: Bangla text classification. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(5), 1–21.

Rajasekar, M., & Geetha, A. (2022). Emotion identification from tamil song lyrics using machine learning algorithms. *International Conference on Advanced Communication and Intelligent Systems*, 324–338.

Rajput, A. (2020). Natural language processing, sentiment analysis, and clinical analytics. In *Innovation in health informatics* (pp. 79–97). Elsevier.

Reily, S. A. (2000). Introduction: Brazilian musics, brazilian identities. *British Journal of Ethnomusicology*, *9*(1), 1–10. https://doi.org/10.1080/09681220008567289

Ribeiro, M. C. (2003). Uma história de regressos: Império, guerra colonial e pós-colonialismo. *Oficina do CES*, *188*, 1–40.

Rizzi, K. R., Maglia, C., Paes, L., & Kanter, M. (2011). O brasil na áfrica (2003-2010): Política desenvolvimento e comércio. *Conjuntura Austral*, *2*(5), ág–61.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Romano, A. (2018). How k-pop became a global phenomenon. *Vox*. https://www.vox.com/culture/2018/2/16/16915672/what-is-kpop-history-explained

Rosebaugh, C., & Shamir, L. (2022). Data science approach to compare the lyrics of popular music artists. *Unisia*, 1–26.

Sasaki, S., Yoshii, K., Nakano, T., Goto, M., & Morishima, S. (2014). Lyricsradar: A lyrics retrieval system based on latent topics of lyrics. *Ismir*, 585–590.

Schneider, A. M. (2021). *Amnesty in brazil: Recompense after repression, 1895-2010*. University of Pittsburgh Press.

Sgard, J. (2003). Hyperinflation and the reconstruction of a national money: Argentina and brazil, 1990-2002.

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222.

Sharma, H., Gupta, S., Sharma, Y., & Purwar, A. (2020). A new model for emotion prediction in music. *2020 6th International Conference on Signal Processing and Communication (ICSC)*, 156–161.

Shtromberg, E. (2021). Contemporary art in brazil, 1960s and 1970s: Forging the "new". In *Oxford research encyclopedia of latin american history*.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*(3), 853–860.

Sievert, C., & Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 63–70.

Silva, C., & Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, *3*, 1661–1666.

Sirisuriya, D. S., et al. (2015a). A comparative study on web scraping.

Sirisuriya, D. S., et al. (2015b). A comparative study on web scraping.

Souza, R. R., & Café, L. M. A. (2018). Análise de sentimento aplicada ao estudo de letras de música. *Informação & Sociedade*, *28*(3).

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). OCTIS: Comparing and optimizing topic models is simple! *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 263–270. https://doi.org/10.18653/v1/2021.eacl-demos.31

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Vieira, C. C., Loures, T. C., de Melo, P. O. S. V., & Assunção, R. M. (2019). The times they are a-changin' (or not): Song lyrics analysis over the years. In J. A. F. dos Santos & D. C. Muchaluat-Saade (Eds.), *Proceedings of the 25th brazillian symposium* *on multimedia and the web, webmedia 2019, rio de janeiro, brazil, october 29 - november 01, 2019* (pp. 237–244). ACM. https://doi.org/10.1145/3323503.3360304

Voutilainen, A. (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, 219–232.

Wall, K., Correia, R. B., & Gouveia, R. (2018). Atitudes face às licenças parentais em portugal. *Sociologia, Problemas e Práticas*, (90). https://doi.org/10.7458/spp20199015525

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in nlp. *COLING 1992 volume 4: The 14th international conference on computational linguistics*.

Whitburn, J. (1996). The billboard book of top 40 hits new york.

Witten, I. H. (2004). Text mining.

APPENDIX A

Summary of some of the Literature found on sentiment analysis and topic modeling

Reference	Article Label	Algorithm / Model / Technique	Target idioms	Key Takeaways
Jo and Kim, 2023	sentiment analysis over time	Structural Topic Modeling (STM)	Korean	The studied period was from the 1990 to 2019. The following negative topics showed significant negative correlations with time: "Being left alone, longing in one's dreams", "Memories about an object you miss on a rainy day", "Longing, heartache experienced over years", "Memories of pain and loving", "Criticizing the world, comparing dreams and reality" The topic "A description of an end, or anger" showed positive correlation with time The following negative topics showed significant negative correlations with time: "Excitement and joy for each other", "Beautiful scenery (flowers, people, streets, etc.), "Morning, tomorrow, a description of a dream", "Sea, waves, summer", "The feeling of falling in love". The following positive topics showed significant positive correlations with time: "Feeling of attraction at the moment", "Action to express or obtain one's heart", "A cheery description of reason for attraction", "A confession of love", "Flowers, spring, sunshine", "Praise for music", "Heart flutter, and other mood (love, eros, etc.)", "Landscape (star, sky, light, city)" "Good" and "Pretty" adjectives increased with time.
de Lucio and Palomeque, 2022	sentiment analysis over decades	Valence Aware Dictionary for Sentiment Reasoning (VADER)	English	When the macroeconomic situation worsens and when weekly insurance claims and is not signifcant for infation, happier and more positive songs become more popular, facing a decrease in songs with negative lyrics, with the assumption that is for people to compensate for the moment circumstances.
Vieira et al., 2019	sentiment analysis over time	word2vec and entropy	English[1]	1. Found rise in the term "supermodel" in the 2000s and in the term "quarrel" in the 1960s. 2. The term "love" shows a high frequency among all decades considered (from 1960s to 2010s). 3. The decade of the 1960s shows particularities in the terms, comparing with the remain decades.
Napier and Shamir, 2018	sentiment analysis over time	IBM Watson Tone Analyzer (Support Vector Machine(SVM))	English[2]	Negative emotions as Anger, Disgust, Fear, Sadness, and Conscientiousness have increased significantly since the 1950s up to 2018, while happy and positive sentiments as Joy, Confidence, and Openness expressed in pop song lyrics have declined.
Brand et al., 2019	sentiment analysis over time	Linguistic Inquiry and Word Count (LIWC) Bayesian, aggregated binomial, multilevel models	English	1. Songs with negative lyrics are more popular 2. There is an increase in songs with negative emotions and a decrease in the ones with positive emotional content 3. The term "love" shows an evident decrease from the middle on the 80's until the 2000's, appearing to increase again from the 2010's 4. Demonstrated a fast increase of the term "hate", however not as rapid as the term "love"
K. Choi et al., 2015	topic modeling	Latent Dirichlet Allocation (LDA)	*not mentioned*	1. This study relevants 6 most popular topics: "Heartache", "Sex", "Parents", "Religion", "Drugs" and "War" 2. The topics consisted in the following terms: "Heartache": relationship, break, feeling "Sex": sex, sexual, girl "Parents": child, father, mother "Religion": god, Christian, religion "Drugs": drug, addiction, heroin "War": war, fight, soldier

APPENDIX B

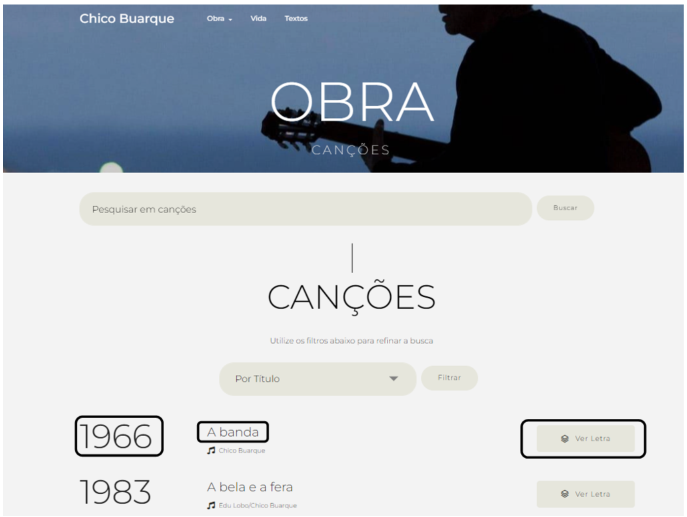
Summary of some of the Literature found on sentiment analysis and topic modeling (cont.)

Reference	Article Label	Algorithm / Model / Technique	Target idioms	Key Takeaways
Rosebaugh and Shamir, 2022	topic modeling	Udat (using deep recurrent neural network (RNN))	English	Studied the sentiments expressed in the lyrics, the diversity in the selection of words, the frequency of gender-related words, and the distribution of the sounds of the words show differences between popular music artists. The analysis also shows a correlation between the easiness of readability and the positivity of the sentiments expressed in the lyrics. 1. Highest frequency of very positive sentences was found in George Harrison's songs, as approximately 4% of the sentences in his songs expressed positive emotions. 2. The positivity was also found in songs by John Lennon, Eric Clapton, Commodores, The Doors, and Queen. 3. It was found that Billy Joel, Boston and Pink Floyd rarely express a positive sentiment in their lyrics, demonstrating a more negative, heavy and darker essence, focusing on political and social topics 4. John Lennon's lyrics (from post-Beatles phase) are also highlighted to incorporate messages on a political and social issues nature, although still keeping a positive tone and emotions conveying in his messages
Pettijohn and Sacco Jr, 2009	topic modeling	Linguistic Inquiry and Word Count (LIWC)	English[0]	Studying the trends on lyrics during more threatening and struggling social and economic times, the authors found a rise in popularity in songs with: 1. more words and with more future references 2. social motivation, sports and social processes (as friendship and talking) topics 3. more personal pronouns
Sasaki et al., 2014	topic modeling	Latent Dirichlet Allocation (LDA) LyricsRadar for visualization	Japanese and english	Use of LyricsRadar tool for topics and detected the following: topic 1: ya, shit, nigga topic 2: baby, yeah, love topic 3: heart, world, life topic 4: bum, clap, jet topic 5: white, Christmas, nice
Cole, 1971	topic modeling	*not mentioned*	English[0]	The authors found that there were more unhappy songs (44%) compared to happy ones (39%) over the entire period of the sixties. The prevalence of happy songs decreased in the second half of the decade. Love and sex were the most common themes (71%), with a constant decrease during the decade. Social protest themes emerged in the latter half of the decade, accounting for 10% of the lyrics. Religious and violent themes were less common, each appearing in about 12% of the songs. References to drugs were not present in the analyzed lyrics.

[0] From Billboard Hot charts, where the majority of the songs are in English.

APPENDIX C

Features extracted from Chico Buarque's official website



OBRA

CANÇÕES

Pesquisar em canções

CANÇÕES

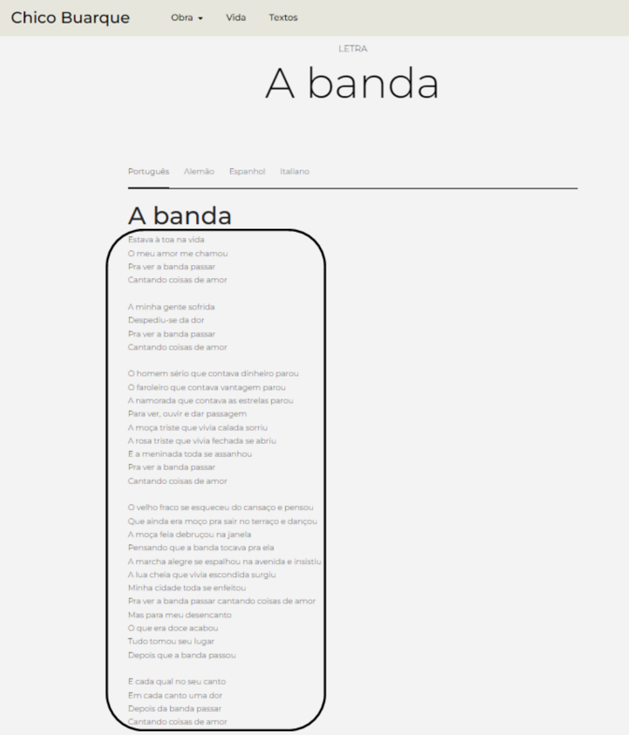
Utilize os filtros abaixo para refinar a busca

Por Título

1966	A banda Chico Buarque	Ver Letra
1983	A bela e a fera Edu Lobo/Chico Buarque	Ver Letra

APPENDIX D

Features extracted from Chico Buarque's official website (cont.)



LETRA

A banda

A banda

Estava à toa na vida
O meu amor me chamou
Pra ver a banda passar
Cantando coisas de amor

A minha gente sofrida
Despediu-se da dor
Pra ver a banda passar
Cantando coisas de amor

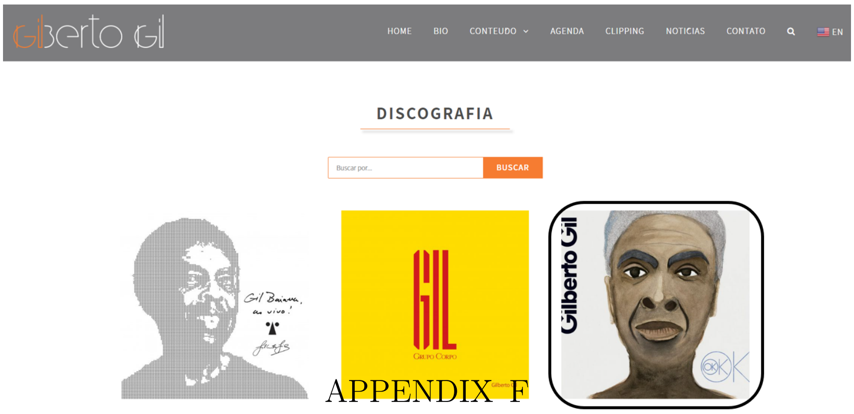
O homem sério que contava dinheiro parou
O faroleiro que contava vantagem parou
A namorada que contava as estrelas parou
Para ver, ouvir e dar passagem
A moça triste que vivia calada sorriu
A rosa triste que vivia fechada se abriu
E a meninada toda se assanhou
Pra ver a banda passar
Cantando coisas de amor

O velho fraco se esqueceu do cansaço e pensou
Que ainda era moço pra sair no terraço e dançou
A moça feia debruçou na janela
Pensando que a banda tocava pra ela
A marcha alegre se espalhou na avenida e insistiu
A lua cheia que vivia escondida surgiu
Minha cidade toda se enfeitou
Pra ver a banda passar cantando coisas de amor
Mas para meu desencanto
O que era doce acabou
Tudo tomou seu lugar
Depois que a banda passou

E cada qual no seu canto
Em cada canto uma dor
Depois da banda passar
Cantando coisas de amor

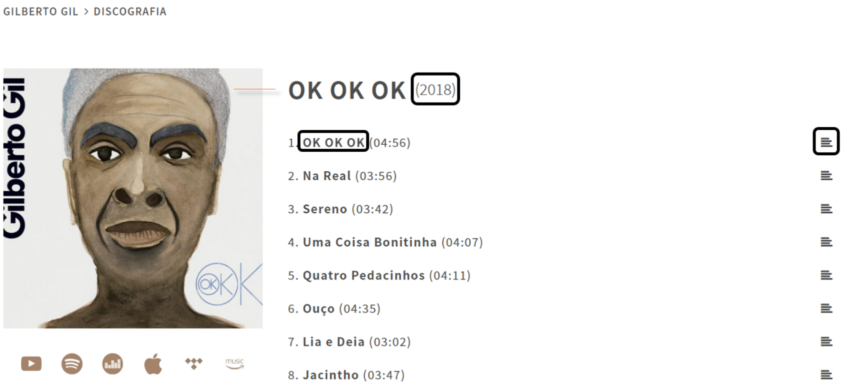
APPENDIX E

Features extracted from Gilberto Gil's official website



APPENDIX F

Features extracted from Gilberto Gil's official website (cont.)

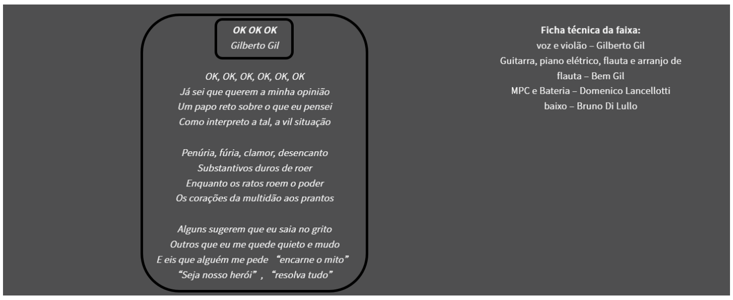


OK OK OK (2018)

1. OK OK OK (04:56)
2. Na Real (03:56)
3. Sereno (03:42)
4. Uma Coisa Bonitinha (04:07)
5. Quatro Pedacinhos (04:11)
6. Ouço (04:35)
7. Lia e Deia (03:02)
8. Jacintho (03:47)

APPENDIX G

Features extracted from Gilberto Gil's official website (cont.)



OK OK OK
Gilberto Gil

OK, OK, OK, OK, OK, OK
Já sei que querem a minha opinião
Um papo reto sobre o que eu pensei
Como interpreto a tal, a vil situação

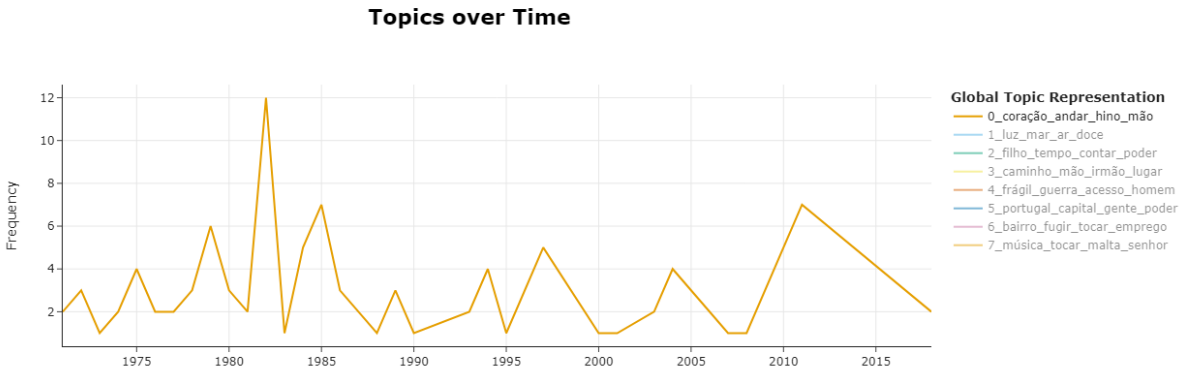
Penúria, fúria, clamor, desencanto
Substantivos duros de roer
Enquanto os ratos roem o poder
Os corações da multidão aos prantos

Alguns sugerem que eu saia no grito
Outros que eu me quede quieto e mudo
E eis que alguém me pede "encarne o mito"
"Seja nosso herói" , "resolva tudo"

Ficha técnica da faixa:
voz e violão – Gilberto Gil
Guitarra, piano elétrico, flauta e arranjo de flauta – Bem Gil
MPC e Bateria – Domenico Lancellotti
baixo – Bruno Di Lullo

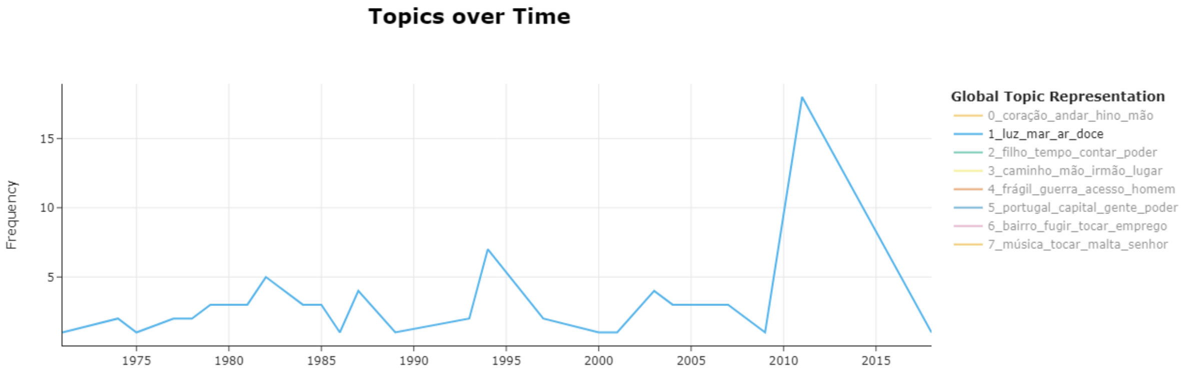
APPENDIX H

Topic 0 Distribution for Portuguese Song Lyrics



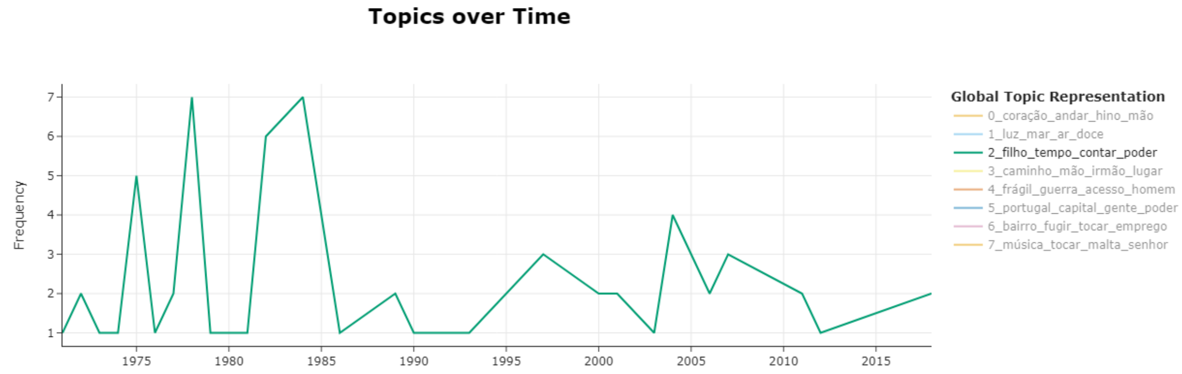
APPENDIX I

Topic 1 Distribution for Portuguese Song Lyrics



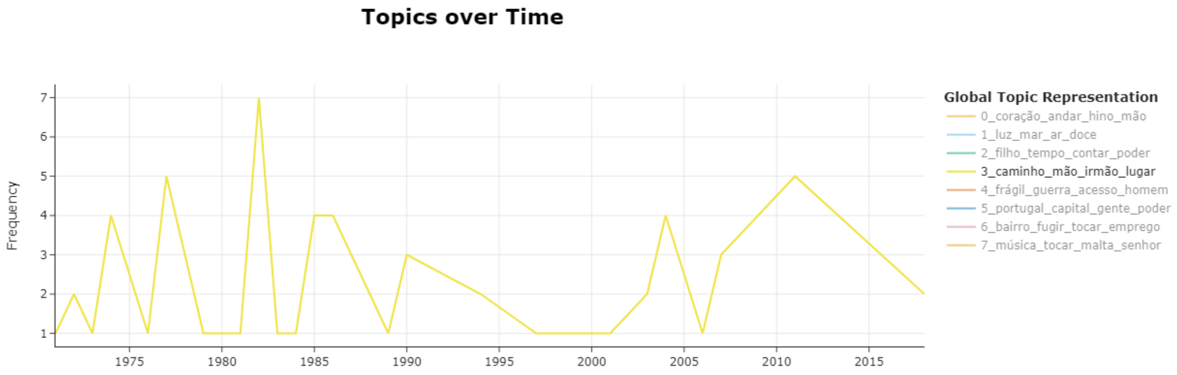
APPENDIX J

Topic 2 Distribution for Portuguese Song Lyrics



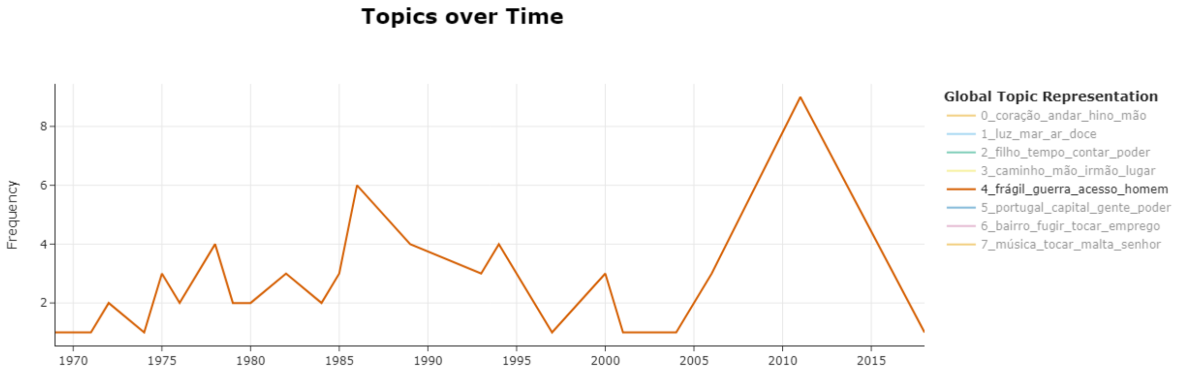
APPENDIX K

Topic 3 Distribution for Portuguese Song Lyrics



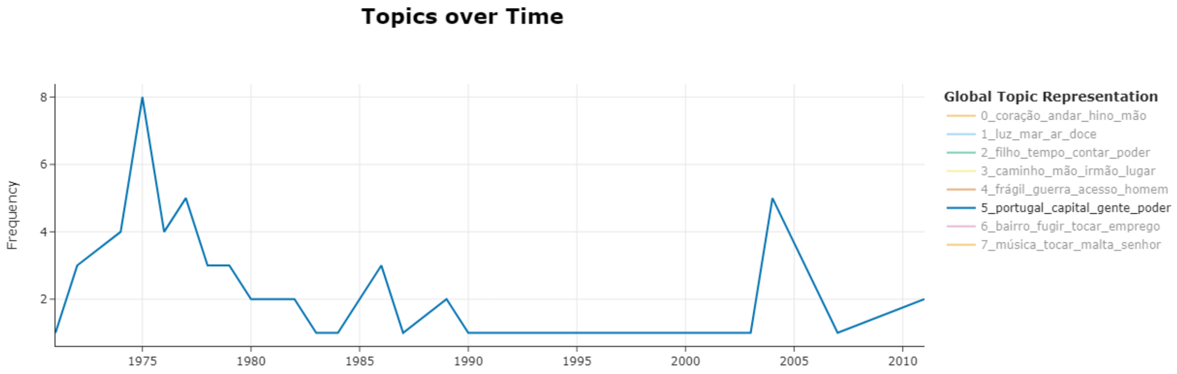
APPENDIX L

Topic 4 Distribution for Portuguese Song Lyrics



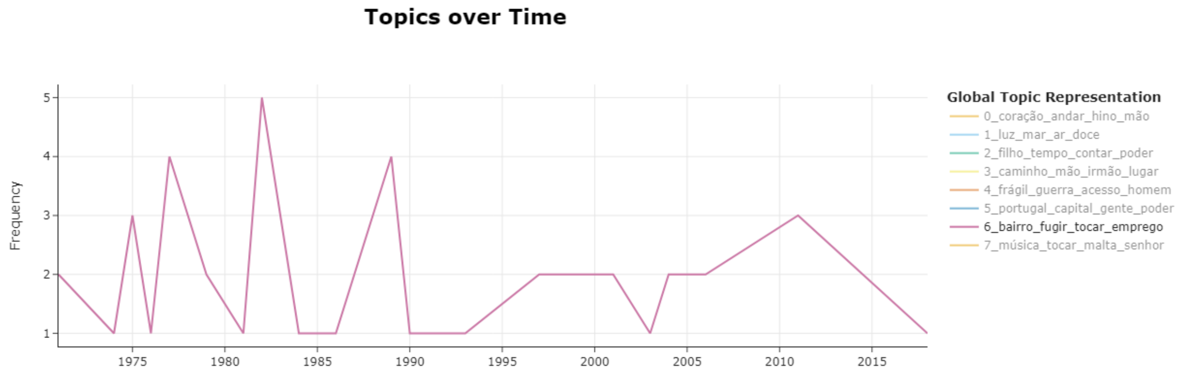
APPENDIX M

Topic 5 Distribution for Portuguese Song Lyrics



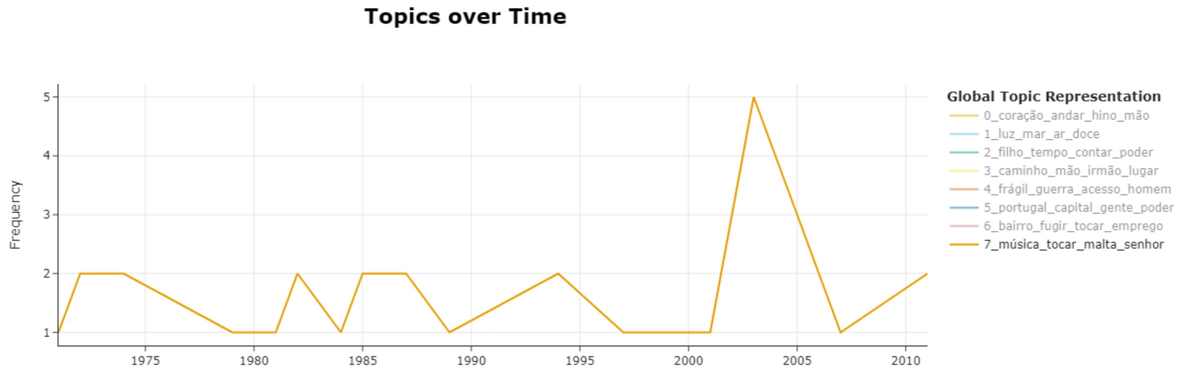
APPENDIX N

Topic 6 Distribution for Portuguese Song Lyrics



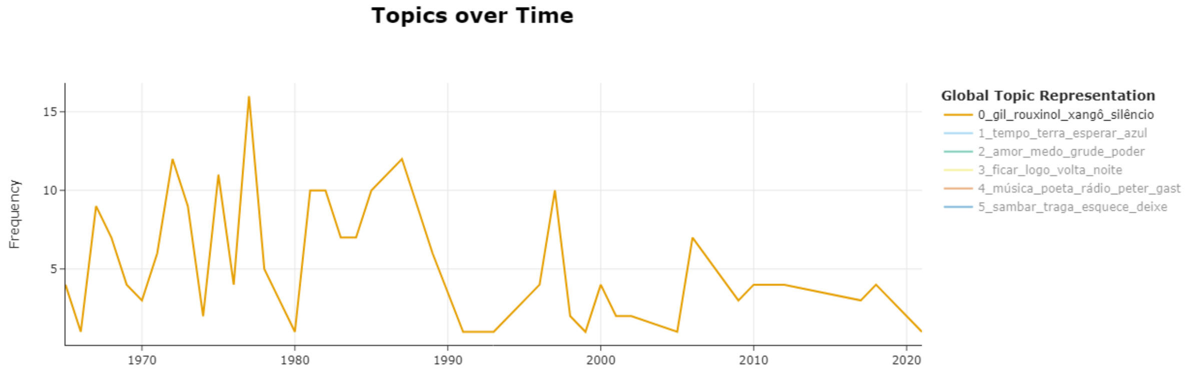
APPENDIX O

Topic 7 Distribution for Portuguese Song Lyrics



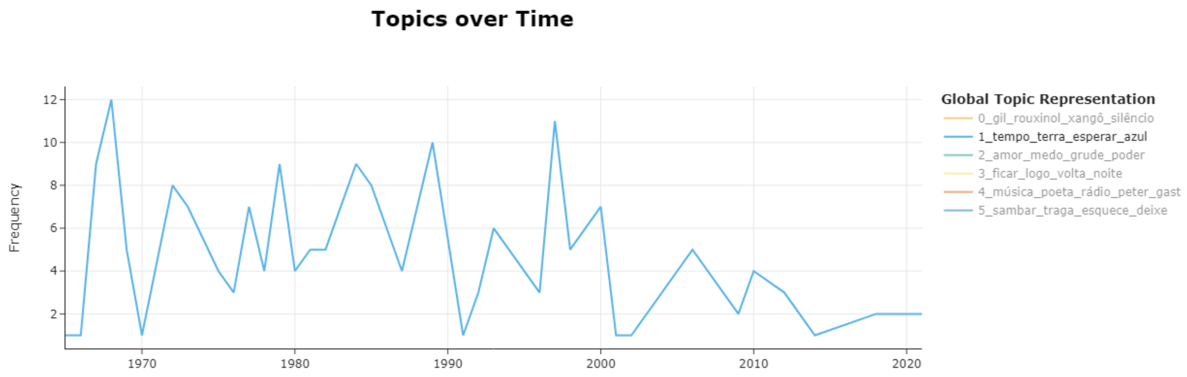
APPENDIX P

Topic 0 Distribution for Brazilian Song Lyrics



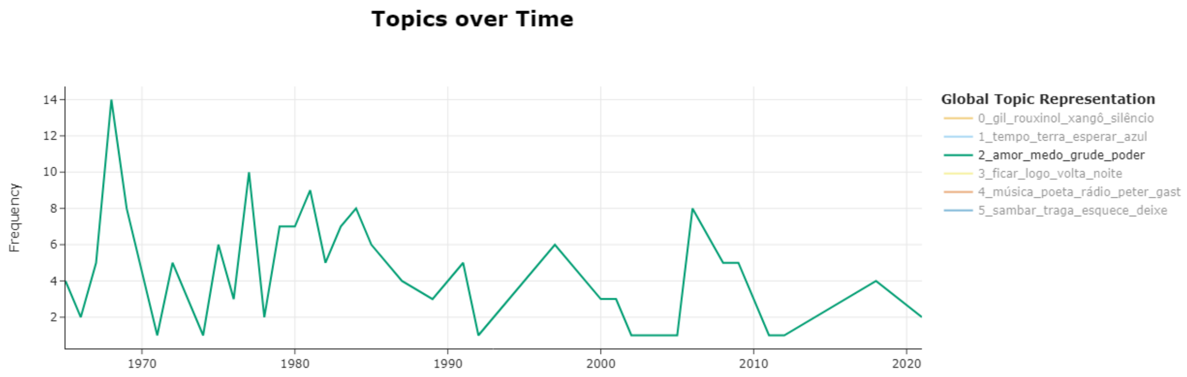
APPENDIX Q

Topic 1 Distribution for Brazilian Song Lyrics



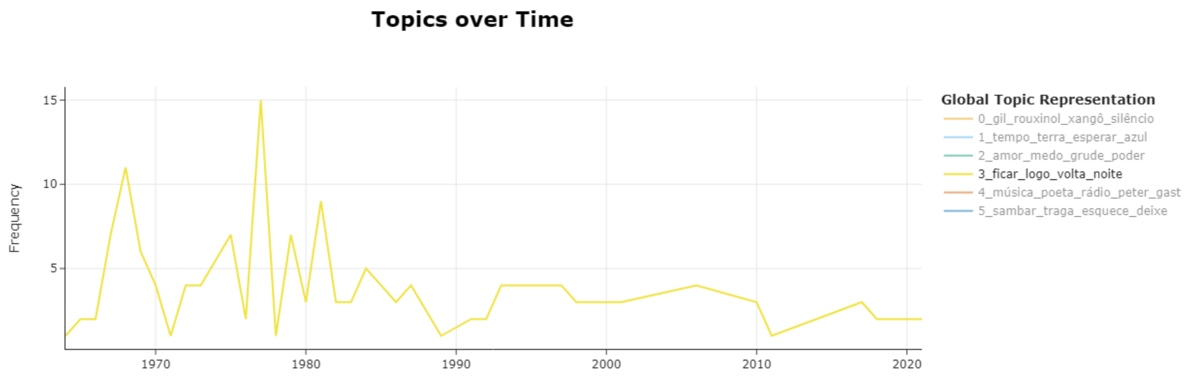
APPENDIX R

Topic 2 Distribution for Brazilian Song Lyrics



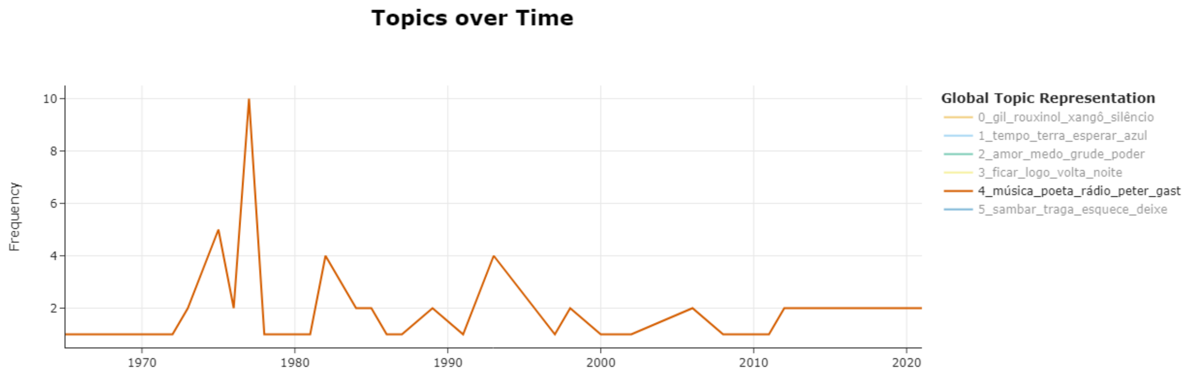
APPENDIX S

Topic 3 Distribution for Brazilian Song Lyrics



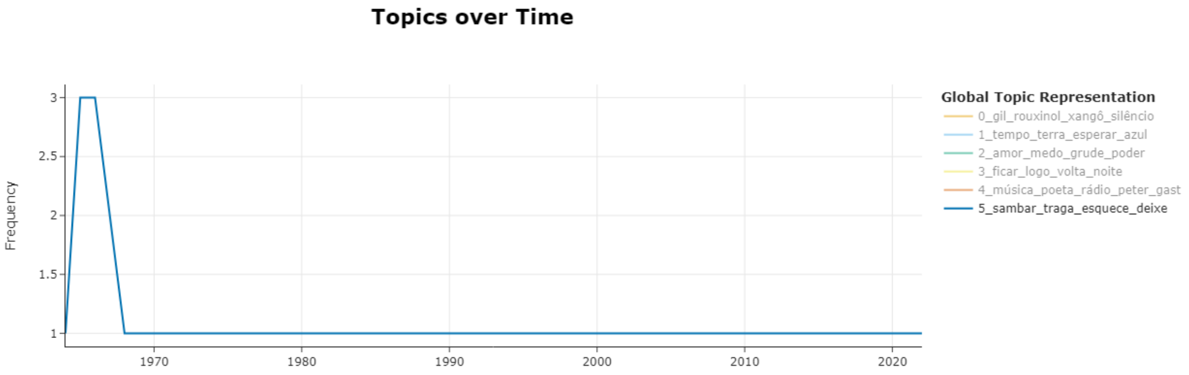
APPENDIX T

Topic 4 Distribution for Brazilian Song Lyrics



APPENDIX U

Topic 5 Distribution for Brazilian Song Lyrics


Topics over Time
Frequency
Global Topic Representation
0_gil_rouxinol_xangô_silêncio
1_tempo_terra_esperar_azul
2_amor_medo_grude_poder
3_ficar_logo_volta_noite
4_música_poeta_rádio_peter_gast
5_sambar_traga_esquece_deixe