Spis treści

1. Introduction

1.1 Problem Statement

Music has been an integral part of human culture for generations. It's been evolving alongside societies, reflecting their creativity, emotions, values and emerging trends. In this day and age music has become more accessible and popular than ever, especially thanks to advancements in technology. Rapidly expanding global audience, growing number of artists and overwhelming number of songs released every day presents us with a great opportunity to investigate more closely its characteristics through the lens of data-driven analyses and machine learning techniques.

Based on the acoustic features that can be extracted from the music tracks, and textual features extracted from song lyrics and information available on Spotify, this paper attempts to understand better complex relationships between different music track characteristics and uncover insights into how music is perceived by listeners, the factors that influence song's popularity, the defining traits of various music genres and how trends shape the evolution of music and its characteristics over the years.

This approach offers a modern, quantitative perspective on music, enabling us to further explore relationships between different musical and textual characteristics.

1.2 Research Objectives

This thesis aims to create a robust and diverse dataset of songs with acoustic features, lyrics and metadata through a well-defined data collection methodology and use advanced data analysis methods and explainable artificial intelligence(XAI) techniques to:

- Analyze relationships between musical and textual characteristics of collected songs
- Examine how song characteristics vary across various genres and what are their defining traits
- Investigate the relationship between lyrical sentiment and acoustic dynamics (e.g., valence and tempo).
- Identify and understand key factors contributing to song's popularity
- Explore music characteristics evolution over the years and shifts in listener's preferences
- Build predictive models to predict various song attributes and use XAI to interpret the models and extract meaningful insights and feature relationships

Additionally, this thesis aims to create a practical and reusable framework for explainable AI (XAI) evaluation, as well as a clear and systematic method for collecting data. These tools are designed to make it easier for future researchers to build on this work, apply the methods to new projects, and explore other areas of music analysis.

1.3 Research Questions

This thesis aims to leverage a diverse dataset of songs collected through a systematic metehodology, consisting of metadata, Spotify's audio features, lyrics and mp3 tracks. With the help of advanced data analysis techniques and explainable artificial intelligence it attempts to achieve following objectives:

1.
2. Explore the relationships between lyrical, acoustic, and metadata features of songs to uncover meaningful patterns and dependencies.
3. Analyze defining characteristics of songs across different genres and investigate how these features contribute to genre identity.

4. Identify key factors that drive song popularity and assess their impact using predictive models and XAI techniques.

5. Develop predictive models to estimate various song attributes (e.g. popularity, explicitness, sentiment) and use XAI to interpret these models, extracting meaningful insights about feature importance and relationships.

1.4 Thesis Structure

This thesis is structured into several chapters to address the research objectives. It begins with an introduction, followed by a discussion of related work, highlighting key studies and their connection to this research. The data chapter explains the sampling method, data collection framework, and sources used.

Next, the methodology chapter details the feature engineering process, describing all extracted features, and introduces the explainable AI (XAI) techniques and other methods applied in the study. This is followed by an extensive data analysis chapter, exploring relationships between features and identifying distinctive traits of each genre.

The experiments chapter presents the conducted experiments and discussion of the results. The thesis concludes with a summary of contributions, a description of the limitations, key findings, and recommendations for future work.

2. Literature Review and Related Work

2.1 Acoustic Features in Music Analysis

“Music Genre Classification Using MFCC, K-NN, and SVM Classifier“

[**music˙genre˙classification˙mfcc**]

This study explores music genre classification using MFCCs and Chroma features on the GTZAN dataset, that consists of 900 tracks across 9 genres. The best-performing model was an SVM with a polynomial kernel, achieving accuracy of 78%. Some genres were identified to have overlapping characteristics which posed problems for the classification model.

The paper highlights the effectiveness of MFCCs and Chroma features for audio-based classification, aligning with this thesis's use of acoustic features. However, unlike this study, the thesis extends the analysis by incorporating lyrical features, metadata, and audio features provided by spotify, addressing limitations in feature diversity and classification accuracy.

“Classifying Music Audio with Timbral and Chroma Features“

[**classifying˙music˙audio**] The paper Ćlassifying Music Audio with Timbral and Chroma Featuresęxamines the use of timbral (e.g., MFCCs) and harmonic (e.g., chroma) features in music classification tasks. The study highlights that MFCCs, which represent timbral qualities, are effective for tasks like **artist identification, achieving 56% accuracy in a 20-way classification task**. When combined with beat-synchronous chroma features, which capture harmonic and melodic information, the model's accuracy improved to 59%. This demonstrates the importance of leveraging complementary audio features for more

robust classification. These findings emphasize the potential of combining diverse acoustic features in predictive models, aligning closely with the objectives of this thesis.

2.2 Lyrical Features in Music Analysis

“Using Machine Learning Analysis to Interpret the Relationship Between Music Emotion and Lyric Features“

[**valence˙and˙lyrics**]

This study investigates the relationship between lyrical features and perceived music emotions using 2,372 Chinese pop songs. Lyric features were extracted with LIWC (Linguistic Inquiry and Word Count), and audio features such as MFCCs and chroma were derived using Librosa. The analysis highlighted that lyrical features like the frequency of positive and negative emotion words contributed significantly to predicting the perceived valence of music, whereas audio features were dominant in predicting perceived arousal.

The study utilized Random Forest regression models, demonstrating that combining lyric and audio features improved **predictions of valence, with an R2 of 0.481**, but lyrics had little impact on arousal models. These findings align with this thesis's use of both lyrical and acoustic features for predictive tasks but differ in the application of explainable AI (XAI) techniques like SHAP for feature interpretation. Moreover, this thesis expands this study by incorporating Spotify audio features and metadata for broader analytical capabilities and more diverse research objectives.

“Sentiment Analysis and Lyrics Theme Recognition Using NLP Techniques“

[**du˙2024**]

This paper investigates the relationship between sentiment and themes in music lyrics using Natural Language Processing (NLP) techniques. It applies sentiment analysis and thematic recognition across a diverse dataset, identifying emotional nuances and recurring topics in the lyrics. The analysis identifies correlations between sentiment categories (positive, negative, neutral) and thematic clusters (e.g., love, social justice, personal reflection). The

authors use techniques like Latent Dirichlet Allocation (LDA) for topic modeling and Support Vector Machines (SVM) for sentiment classification.

In the context of this thesis, this study aligns closely with the focus on lyrical analysis, particularly in employing NLP-driven sentiment and thematic classification. However, while this work emphasizes standalone lyric-based analysis, this thesis extends the methodology by integrating Spotify's audio features, acoustic analysis, and metadata.

2.3 Machine Learning in Music Analysis

"Beyond Beats: A Recipe to Song Popularity? A Machine Learning Approach"

[**beyond˙beats**] The paper "Beyond Beats"investigates the predictive power of various machine learning models for song popularity on a dataset of 30,000 songs spanning six genres. It focuses on metadata and audio features fetched from Spotify(*danceability*, *acousticness* etc.), as well as the genre. The best performing model was **Random Forest and achieved 16.31 MAE.** Predictions across all methods remained relatively modest, reflecting the complex and multi-dimensional nature of song popularity. Those results can be greatly improved using additional features. The authors noted that predictive accuracy was constrained by the absence of post-release factors such as marketing, social media reception, and artist reputation.

"Predicting Song Popularity in the Digital Age Through Spotify's Data"

[**predicting˙song˙popularity˙2024**] Similarily to the previous one, the paper "Predicting Song Popularity in the Digital Age Through Spotify's Dataęxplores the relationship between Spotify's audio features and song popularity, using a dataset spanning from 1986 to 2022. The study employed linear regression to predict popularity and **achieved an adjusted R-squared of 0.38**, highlighting the moderate predictive power of these features. The analysis revealed that attributes such as *danceability* and *duration* positively correlate with popularity, while *speechiness* tends to have a negative impact.

3. Data

3.1 Sampling Method

Spotify's catalog contains approximately 100 million songs and over 6000 distinct genres. Collecting a truly random sample from this huge and diverse population presents significant challenges. A purely random sampling approach would likely result in a dataset heavily skewed towards obscure genres and imbalanced release year distribution. This would lead to poor data quality for meeting the research objectives, with sparse meaningful relationships and an overwhelming diversity of genres. Such diversity would make it nearly impossible to extract actionable insights and relationships, as less represented genres and rare tracks would dominate, shifting the focus from more prominent trends and patterns in music.

To address these challenges, a smaller and more organized sample of 3500 songs was selected. The stratified sampling approach was designed to ensure a balanced dataset by focusing on two key factors: genre and release year. The sampling process utilized Spotify's query parametrization tool, which allows for the specification of desired genres and release year ranges for musical tracks. Release years were grouped into 10-year intervals, starting from 1950 and extending to 2020, ensuring the dataset represents different time periods in modern music. The selection of genres was arbitrary and aimed to include a mix of popular and diverse styles, providing a robust representation of mainstream musical styles.

While this approach introduces some degree of selection bias, favoring tracks that Spotify's search algorithm prioritizes, this bias is acceptable given the research objectives. By explicitly focusing on popular genres and balancing across release years, the sample aims to capture relationships and patterns that are representative of mainstream trends and music

evolution. The structured stratification ensures that each genre and time period is adequately represented, providing a comprehensive and interpretable dataset for the analysis.

3.2 Dataset Description

The dataset consists of around 3500 songs. For each song metadata, audio recording, lyrics and spotify audio features were fetched.

3.3 Data Collection Methods

3.3.1 Sources

The data was collected from various sources. Metadata and audio features were fetched from Spotify API. The lyrics were scraped from LetrasMus, MakeItPersonal or Lyrics Fandom, or fetched via the Genius API, depending on the availability. Audio files were downloaded from YouTube and saved as mp3 files.

3.3.2 Methodology

The data collection process was automated for efficiency. The starting point was a list of Spotify playlist URIs. Each playlist was processed sequentially. The script fetched metadata and audio features for each song in the playlist.

Lyrics were then searched based on *artist name* and *title* of each song, using multiple lyrics providers. The system continued to query different sources until it found lyrics in at least one of them. If it failed to retrieve lyrics for the song, it was discarded and wouldn't make it to the final dataset.

Finally, the script searched YouTube for each song and downloaded the first relevant result, saving it to an mp3 file. All data was saved into a CSV file named after the playlist and stored alongside the recordings.

The entire process was parallelized to significantly enhance the speed of data acquisition. It was designed in a robust manner, with careful error handling and ability to stop

the process at any time and pick up where it left off.

Additionally, to optimize performance, the entire process was parallelized, significantly increasing the speed of data acquisition. It was designed with robust error handling, ensuring data correctness and completeness. Moreover, it featured the ability to pause and seamlessly resume the process, continuing exactly where it left off without data loss.

3.3.3 Metadata

Song's information downloaded from Spotify API. The information it includes is:

- **Popularity** - relative measure with values ranging from 0 to 100 describing how popular the song is, estimated mostly based on total number of plays and how recent those plays are
- **Explicitness** - whether or not the song contains explicit lyrics
- **Genre** - main genre of the artist(Spotify does not provide information about genre of each musical track)
- **Album Release Year** - the release year of the album that the song originates from

popularity, *explicitness*, *genre*, and *release year*. Since Spotify does not provide information about the genre of specific songs, the genre extracted is the main genre of the primary artist in the song.

3.3.4 Spotify Audio Features

Those features were fetched from the Spotify API.They describe different acoustic properties songs:

- **Speechiness** - relative measure of spoken words in a track
- **Acousticness** - a confidence measure of whether the song is acoustic
- **Danceability** - a measure of how suitable the song is for dancing. Its based on parameters like tempo, rhythm stability, beat strength etc.

- **Energy** - a measure of perceived intensity of songs. Energetic tracks are usually louder, faster, feel more intense.
- **Loudness** - overall loudness of the track in dB averaged across the entire track
- **Valence** - relative measure describing musical positiveness of a track
- **Instrumentalness** - measure of how likelihood of the track not containing vocals. In this paper since lyrics are mandatory it's used to discard instrumental tracks.
- **Liveness** - probability of the song being recorded during a live performance.
- **Key** - the key of the song, e.g. C#.
- **Mode** - indicates the modality of a track(major / minor)
- **Tempo** - estimated tempo of a track in beats per minute(BPM)
- **Time Signature** - specifies how many beats there are in each bar
- **Duration** - the duration of the track in milliseconds

3.3.5 Lyrics Features

The lyrics serve as a textual representation of the song's thematic, emotional, and linguistic elements. Since they came from various data sources, they had to undergo cleaning procedure in order to remove faulty information and prepare them to be processed by the textual feature extraction class. The process consisted of:

- **Standardization** - lyrics were converted to lowercase
- **Noise Removal** - unnecessary characters, numbers and punctuation, as well as additional comments used by lyrics providers(e.g. 'chorus') were removed
- **Stopwords Filtration** - exclusion of frequently occuring words that carry little information, like 'the' in English

- **Stemming** - words were reduced to their root forms to enhance uniformity and reduce corpus size

This process laid foundation for further extraction of textual features used for exploratory data analysis, statistical inference and training ML models. That process will be explained in later chapter.

3.4 Tools and Libraries Used

Python libraries used to facilitate the data acquisition were:

- *Spotipy* [**spotipy**]- a lightweight python library for Spotify API
- *youtube-dl* [**ytdl**] - a library used to find and download YouTube videos
- *BeautifulSoup* [**beautifulsoup**] - a library used for extracting information from HTML, commonly used for web scraping

4. Methodologies

This chapter explains the methodologies used to address the research objectives in this thesis. It describes the feature extraction and engineering methods, the use of explainable AI methods, model training and optimization approaches, clustering and dimensionality reduction techniques and statistical hypothesis testing methods.

4.1 Feature Engineering

Feature engineering is a critical step in the research process, as it involves transforming raw data acquired using the data collection script into meaningful representations that can be analyzed or used for predictive modeling. This section describes the methodologies used to extract acoustic features from the mp3 files and lyrical features from the lyrics. These features are designed to capture key characteristics of the songs, enabling deeper insights into their patterns and relationships.

4.1.1 Acoustic Features

These features provide a quantitative representation of the audio properties of each song and were extracted directly from the audio files in MP3 format. They describe various aspects of audio signal and provide insights into the rhythm, timbre, harmony and other acoustic properties. The extraction was done using *Librosa* and was automated and parallelized to make it suitable for processing large amounts of data.

MFCC - Mel Frequency Cepstral Coefficients

MFCCs represent the short-term power spectrum of a song on a mel-scale and are widely used for timbre analysis. These coefficients capture the tonal quality of the audio and help differentiate between different instruments and vocal characteristics.

Chroma

Chroma vectors represent the intensity of each pitch class (e.g., C, C#, D, etc.) in the audio. These features provide a harmonic representation of the song and are useful for analyzing chord progressions and harmonic structures.

Spectral Contrast

Spectral contrast measures the difference in amplitude between peaks and valleys in the spectrum. It provides insights into the harmonic and timbral content of a song, particularly useful for distinguishing between smooth and complex textures.

Other Features

Two additional features were extracted:

- **Tempo** - refers to the speed of the song, measured in *Beats Per Minute(BPM)*
- **Zero Crossing Rate(ZCR)** - measures the rate at which the audio signal changes sign. It's commonly used as a measure of noisiness or percussive nature of signal

4.1.2 Lyrical Features

Lyrical features were extracted from the lyrics fetched during the data collection process. They aim to provide a linguistic and semantic representation of the track, capturing their complexity, sentiment and stylistic attributes. The cleaning and extraction process utilized various NLP libraries like *NLTK*, *spaCy* and *TextBlob*, alongside with custom ad-hoc

algorithms. Similarily to acoustic features the implementation allowed for simple and intuitive usage under clear and comprehensible interface, with parallelization of the computation process for increased performance. The features extracted can be grouped as follows:

Basic Linguistic Metrics

- **Unique Word Count** - measures number of unique words in the lyrics, indicating diversity
- **Type-Token Ratio** - a measure of lexical richness: ratio of unique words to total words
- **Word Count** - total number of words, baseline for text size and compexity
- **Noun and Verb Ratios** - proporrtios of nous and vers relative to the total word count

Sentiment and Emotional Tone

- **Sentiment Polarity** - a measure of overall sentiment(positive vs. negative) of the text
- **Sentiment Subjectivity** - represents the degree of subjectivity in the lyrics, attempting to make a distinction between factual and opinionated content
- **VADER Compound** - a sentiment score derived from the VADER tool
- **Sentiment Variability** - standard deviation of sentiment on subsets of lyrics, a metric aiming to capture fluctuations of sentiment throughout the song, highlighting emotional complexity

Stylistic Features

- **Repetition Count** - the frequency of repeated words
- **Rhyme Density** - a measure of how often rhymes occur in the text

Semantic and Complexity Features

- **Semantic Depth** - represents the richness and variety of meaning conveyed by the lyrics
- **Syntactic Complexity** - captures the sophistication of sentence structures
- **Lexical Richness** - quantifies the variety and richness of the vocabulary

Readability and Accessibility

- **Flesch Reading Ease** - indicates how easy the lyrics are to read
- **Gunning Fog** - a metric that estimates the years of education required to understand the text
- **Dale Chall** - a metric that accounts for familiar and unfamiliar words in the text

Contextual Information

In process of feature extraction the **language** of lyrics was also identified using *langdetect* library that uses a classification model to make predictions based on n-grams extracted from the text. The identified language was also used in the cleaning process, to identify which stemmer and stopwords language to use.

TF-IDF (Term Frequency - Inverse Document Frequency)

TF-IDF[**tfidf**] is a measure that can quantify the relevance of tokens in a document amongst a collection of documents. In the context of this study, it determines how important a word in a song's lyrics is compared to all other song's lyrics in the dataset. It's used to highlight words that are unique or meaningful while giving less importance to very common words like 'the' or 'and'.

It can be broken down into two parts:

- **TF - Term Frequency** - Measures the frequency of a term within a document:

$$TF(w, d) = \frac{f_{w,d}}{N_d}$$

  where:

  - $f_{w,d}$: The number of times the word w appears in the document d.
  - N_d: The total number of words in the document d.

- **IDF - Inverse Document Frequency** - Measures the rarity of a term across a collection of documents:

$$IDF(w) = \log \frac{N}{1 + n_w}$$

  where:

  - N: The total number of documents in the corpus.
  - n_w: The number of documents containing the word w.

Empath

Empath[**empath**] is a semantic analysis tool designed for extracting semantic, thematic and emotional features from text by analyzing words based on predefined categories. It classifies words into over 200 built-in categories, such as *sadness*, *food* and *music*. It identifies relationships between words and these categories, offering a higher-level representation of text content beyond simple word frequency.

In this study, Empath was used to analyze song lyrics, allowing for extraction of thematic and emotional content. These features were used to complement other lyrical reprersentations, such as TF-IDF and LDA. In contrary to those methods however, empath provided a broader contextual understanding of lyrics by summarizing them into measningful categories. This approach also enhanced the interpretability of topics identified by LDA, as empath features offered an additional layer of semantic insight, making it easier to characterize and describe those topics.

Moreover, Empath allows for a more nuanced exploration of lyrical content across different genres. By analyzing trends in emotional categories like sadness or anger, and themes such as party or work, it helps to identify meaningful variations between genres or time periods.

By combining Empath with traditional feature extraction methods, this study benefits from a richer, more interpretable representation of lyrical content.

4.2 Explainable AI Methods

Explainable AI (XAI) techniques provide insights into the decision-making processes of ML models, making it possible to understand the complex relationships captured within the training data. By bridging the gap between the pattern-recognition capabilities of these models and their practical applications, XAI enables transparency and improves the interpretability of results.

In this study this methodology was applied in various experiments to understand how specific variables influence others, with aim of uncovering rrelationships in the data and validating hypotheses. A deeper understanding of factors driving model predictions ensured that the results were both reliable and meaningfully adressed the research objectives. The techniques employed in this paper are:

SHAP (SHapley Additive exPlanations)

SHAP[**shap**] values offer a detailed breakdown of how individual features contribute to each prediction. It uses game theory principles to compute the importance of each feature for a given output, providing both global insights on general feature importance and local explainations, showing how diffrenet features contributed to specific predictions. In this study SHAP values were used in order to explain the predictive process of trained Catboost models, allowing for identification of key features for specific prediction task and visualization of relationships.

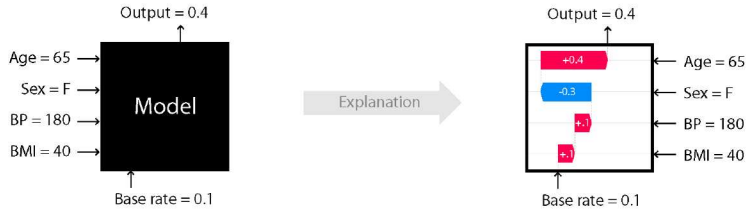

Figure 4.1: SHAP[**shap**]

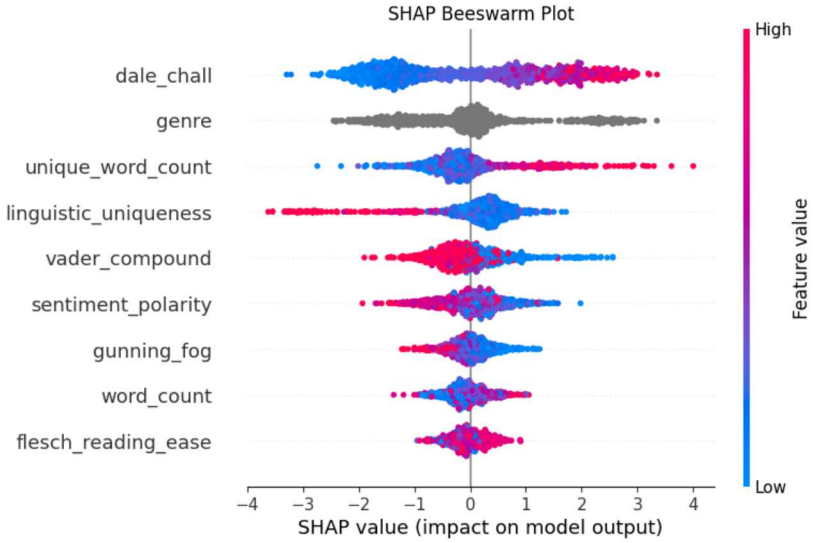

Figure 4.2: Example SHAP beeswarm plot showing impact of some lyrical features on the classifier of *explicitness*

Machine Learning Models

In this study, CatBoost, a widely recognized gradient boosting algorithm known for its high predictive accuracy and efficiency, was employed to build classification and regression models. These models utilized a combination of acoustic, lyrical and metadata features in order to predict the variable of interest, leveraging Catboost's strengths in handling diverse and complex datasets. The result models were then subjected to SHAP analysis, in order to understand their decision process and understand the interactions between features and the target variable. Catboost was chosen for its flexibility, ease of use, robust performance,

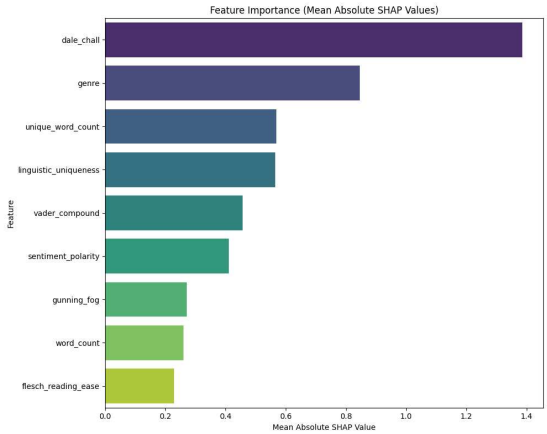

Figure 4.3: Example SHAP feature importance plot showing impact of some lyrical features on the classifier of *explicitness*

compatibility with SHAP and built-in support for categorical features, which eliminated the need for extensive preprocessing.

In order to further optimize the performance of these models, the hyperparameter tuning library *Optuna* was used. Its efficient optimization framework allowed for systematic exploration of different sets of hyperparameter configurations, ensuring the model achieved optimal performance while avoiding overfitting.

To address the challenges commonly encountered when training ML models on complex datasets, following techniques were employed:

- **Cross-validation** - cross-validation was used to reduce the risk of overfitting and provide reliable performance metrics. By dividing the data into multiple folds, the model was iteratively trained and validated on different subsets, therefore ensuring robust evaluation across the dataset and improved model's reliability, at the cost of increased computational time.

- **Class Weights** - to handle class imbalance in target variable labels, CatBoost offers a built-in capability to assign different penalties for misclassifications of specific classes. This adjustment improves model's ability to make accurate predictions across all classes, instead of favouring the majority class.

- **Out of sample evaluation** - model's performance was assessed on a separate test

dataset that was excluded from the training. This step provided a reliable measure of model's ability to generalize on unseen data and ensured evaluation metrics reflected its real predictive performance.

4.3 Dimensionality Reduction - Principal Component Analysis (PCA)

Principal Component Analysis (PCA) reduces the number of features in large datasets by transforming them into principal components that retain most of the original information. It achieves this by converting potentially correlated variables into a smaller set of less correlated variables, called principal components, in a way that preserves as much of the original variance as possible[**pca**]. PCA is often employed to reduce dataset dimensionality and improve generalization by reducing noise and redundancy in the data.

In this study, PCA was applied to the TF-IDF vectors derived from song lyrics. TF-IDF vectors are typically high-dimensional, with thousands of features representing individual terms across the corpus. Such high-dimensional data can pose challenges, including increased computational complexity and a higher risk of overfitting in machine learning models.

The use of PCA on these vectors reduced their dimensionality while preserving as many significant patterns from the original vectors as possible. This process improved computational efficiency and the interpretability of the data, which is particularly important in the context of Explainable AI (XAI) methodologies.

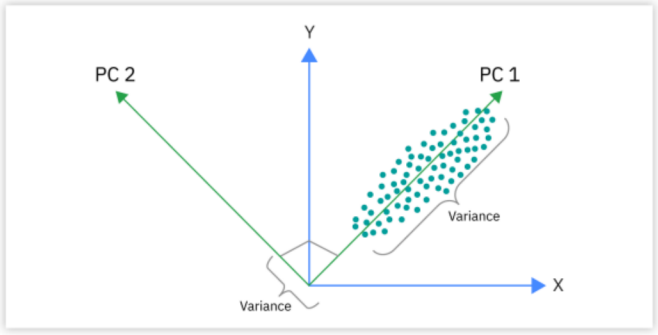

Figure 4.4: A scatterplot showing the relationship between PC1 and PC2 when PCA is applied to a dataset. PC1 and PC2 axis are perpendicular to each other.[**pca**]

4.4 Topic Modelling - Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)[**lda**] is a generative probabilistic model designed to uncover latent topics within a collection of discrete data, such as text corpus. It represents each document as a mixture of topics, where each topic is characterized by a distribution over words. In this study, LDA was applied to the lyrics dataset to identify prevalent themes and topics across different songs. By representing each song as a distribution over topics, LDA provided insights into the thematic content of the lyrics, allowing for the analysis of how these themes correlate with acoustic features and other song attributes.

In this study, LDA was applied on the lyrics in order to identify commonly ocurrirng topics in songs. Each song's lyrics were represented as a combination of topics, and the most representative words for each topic were extracted. This provided insights into the thematic content of lyrics, enabling further exploration of relationships between lyrical topics and other features, like popularity orr acoustic properties.

By combining the topics derived with LDA with additional features, such as acoustic parameters, the study aimed to analyze the interplay between a song's musical and lyrical components.

4.5 Statistical Methods

Statistical methods provided a foundation for analyzing complex relationships between features and a framework for descriptive data analysis and hypothesis testing.

4.5.1 Pearson Correlation

Pearson Corrrelation is a statistical measure used to quantify linear relationship between two continuous variables. The resulting coefficient ranges from -1 to 1 and is computed in the following way:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

- x_i and y_i: The data points for the two variables.
- $\bar{x}$ and $\bar{y}$: The mean values of the variables.

Interpretation:

- An r value close to 1 indicates very strong positive linear relationship
- An r value close to -1 indicates very strong negative linear relationship
- An r value close to 0 indicates little to no relationship

4.5.2 Bootstrap Testing

Bootstrap testing was used to verify hypotheses specified in the research objectives. It's a resampling-based statistical technique that estimates the variability of a statistic(e.g. mean or median) by repeatedly sampling its values from the dataset with replacement. It's highly versatile since it doesn't rely on strong distributional assumptions.

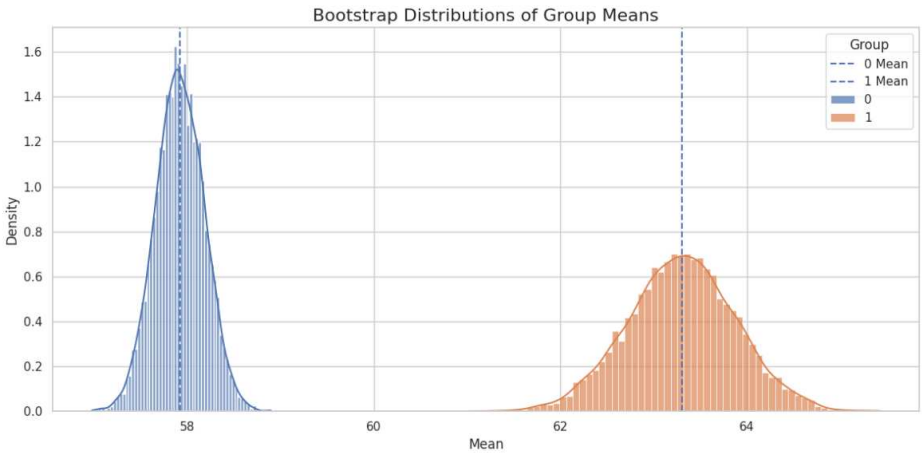

Figure 4.5: Illustration of bootstrap resampling: The distribution of the statistic (e.g., mean) of the target variable for two different samples (called groups). This demonstrates the variability of the statistic across resampled datasets.

5. Exploratory Data Analysis

5.1 Spotify Features

Correlation Heatmap

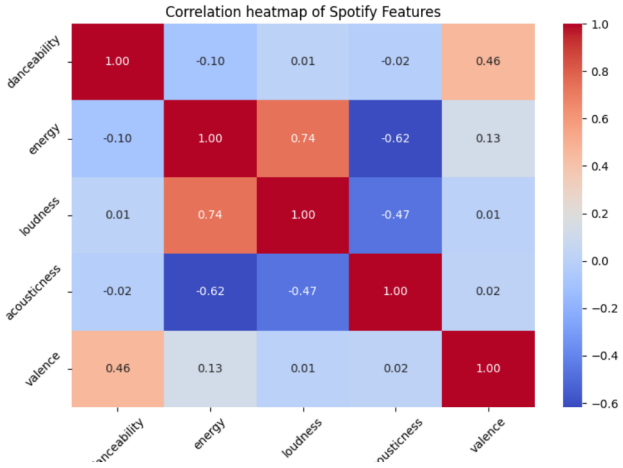

Figure 5.1: Pearson's Correlation Heatmap of Spotify Features

Hierarchical Clustering

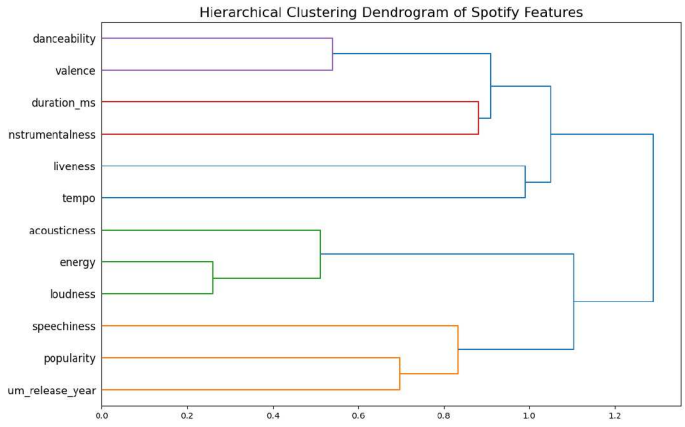

Figure 5.2: Hierarchical Clustering of Spotify Features

Observations

Spotify audio features show high level of correlation between each other, especially:

- *Energy* and *Loudness*: A high correlation(0.74) suggest that energetic songs are typically louder.
- *Danceability* and *Valence*: high correlation between them indicates that songs perceived as positive and happy are usually more danceable.
- *Energy* and *Acousticness*: there is a strong negative correlation between those features, suggesting that high-energy tracks are less likely to have acoustic elements.
- The dendrogram5.14 shows relations between features and allows us to compare which features correlate with each other. In addition to the relationships seen in the correlation heatmap, we observe that *popularity* appears to have a weak correlation with *release year* and *speechiness*.

5.2 Lyrical Features

Correlation Heatmap

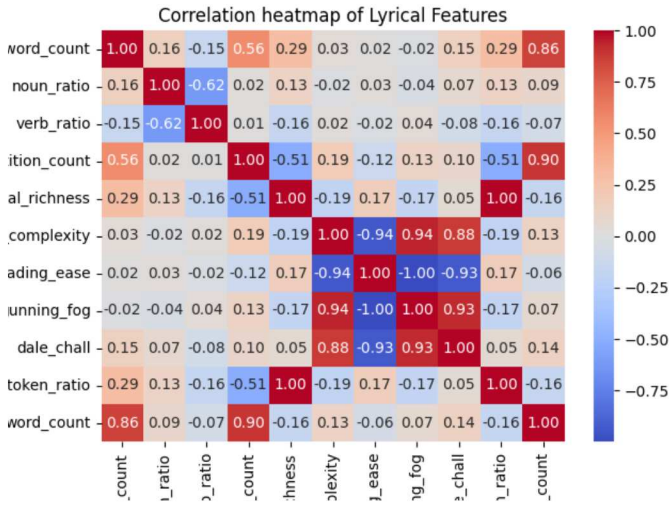

Figure 5.3: Pearson's Correlation Heatmap of Lyrical Features

Hierarchical Clustering

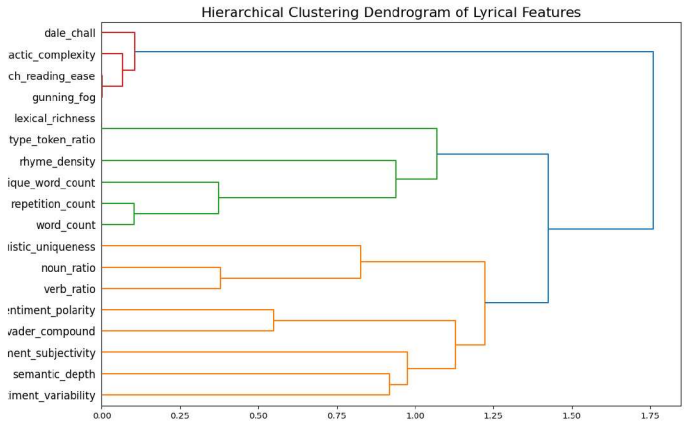

Figure 5.4: Hierarchical Clustering of Lyrical Features

Observations

- *Flesch Reading Ease*, *Gunning Fog* and *Dale Chall* scores exhibit strong positive correlations, highlighting their shared focus on measuring lyrical complexity.
- *Lexical Richness* correlates with *Word Count*, indicating that more lexically rrich lyrics tend to have greater variety of words.
- On the dendrogram we can distinguish three major clusters of features:
  - **Lyrical Complexity Metrics** - *Flesch Reading Ease*, *Gunning Fog*, *Dale Chall* and *Syntactic Complexity* quantify how difficult and complex the lyrics are.
  - **Lexical Features** - features such as *Type-Token Ratio*, *Lexical Richness* and *Unique Word Count* form a cohesive cluster.
  - **Sentiment-Related Features** - it includes *Sentiment Polarity*, *VADER Compound* and *Semantic Depth*, which reflect emotional aspects of the lyrics.

- Sentiment-related features are relatively closely grouped, suggesting a high level of interdependence

5.3 Audio Features

Correlation Heatmap

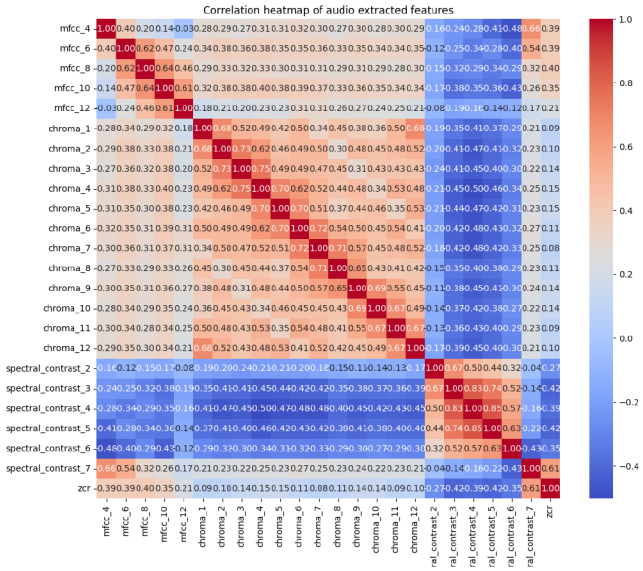

Figure 5.5: Pearson's Correlation Heatmap of Audio Features

Hierarchical Clustering

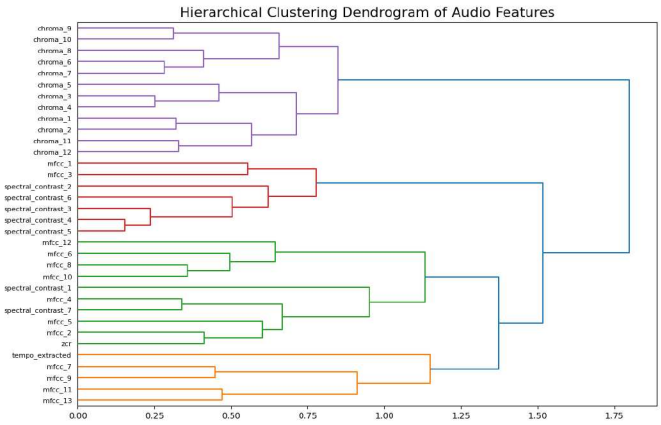

Figure 5.6: Hierarchical Clustering of Audio Features

Observations

- *MFCC Features* show strong correlations among themselves(e.g. *mfcc_4*, *mfcc_6*, *mfcc_8*, etc.).
- Similarily *chroma* features are highly correlated with each other, indicating that they capture similar aspects of tonal energy distribution.
- *spectral_contrast* features exhibit relatively small correlations with *mfcc* and *chroma* features, suggesting that they capture different characteristics of audio
- All *chroma* features are grouped together in the same cluster on the dendrogram, supporting the conclusion drawn from the heatmap about strong correlations between them
- *zcr* and *tempo_extracted* are relatively independent

- A certain degree of overlap between *mfcc* and *spectral_contrast* features is observed in the dendrogram, indicating potential shared information in capturing specific audio properties.

5.4 Empath Features

Correlation Heatmap

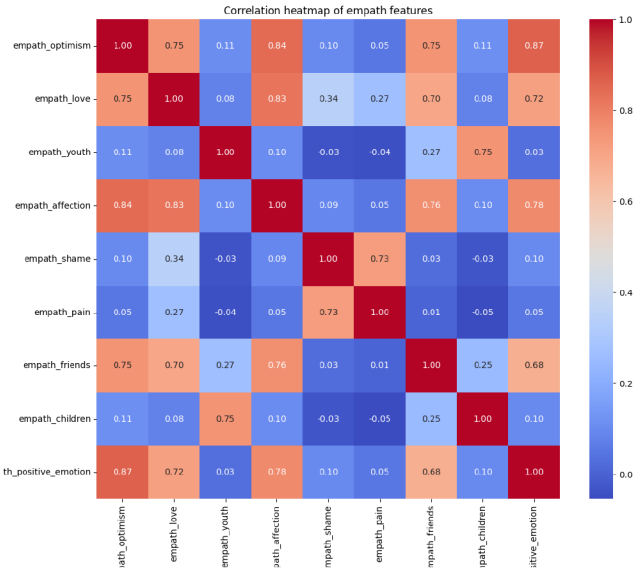

Figure 5.7: Pearson's Correlation Heatmap of Empath Features

Hierarchical Clustering

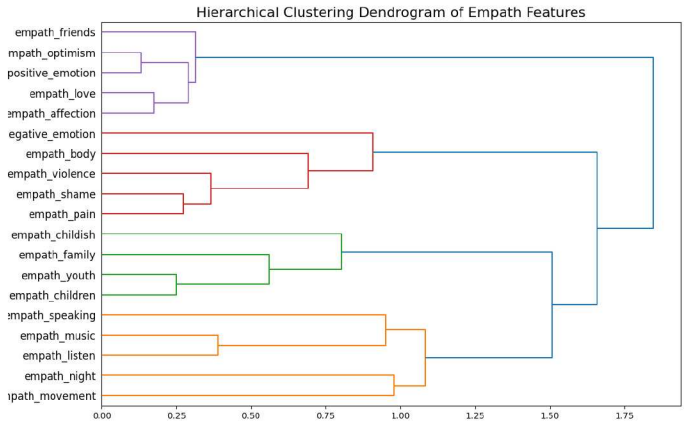

Figure 5.8: Hierarchical Clustering of Empath Features

Observations

- **High Correlations Reflect Logical Groupings**: high correlation between features such as *empath_optimism*, *empath_love* and *empath_affection* align with the intuitive understanding that these aspects are closely related.
- Observed patters in corrrelations confirm that Empath was designed to group semantically related concepts together.
- The visualizations align with the intended design of Empath as a tool for interpretable feature extraction.

5.5 Genre Analysis

5.5.1 Genre Similarity

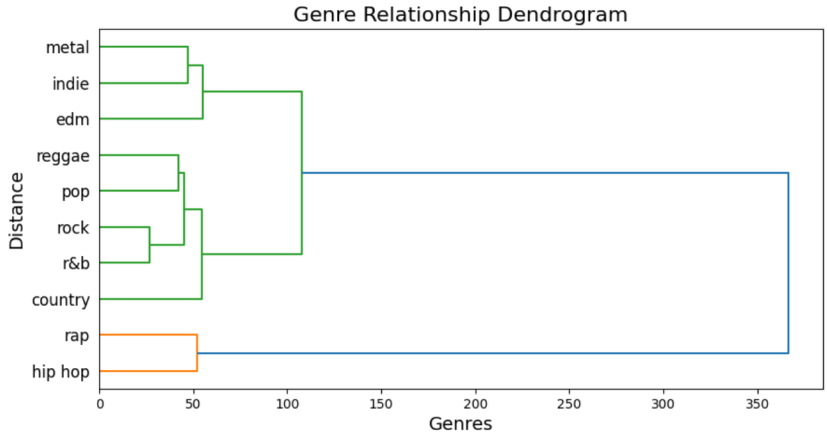

Figure 5.9: Hierarchical Clustering of Genres by lyrical and audio features

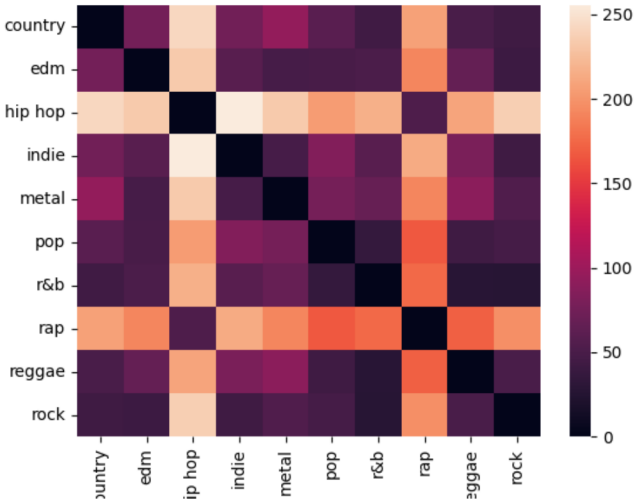

Figure 5.10: Heatmap of euclidean distance of Genres calculated on lyrical and audio features

The dendrogram illustrates hierarchical clustering of genres based on their lyrical and audio features. The clustering aligns with general cultural and musical understandings of these genres. For instance, closely related genres like hip hop and rap are grouped together, indicating that they're highly similar.

The heatmap further supports this observation. The Euclidean distances of *rap* and *hip hop* with other genres stand out as the highest.

Interestingly, genres such as *metal* and *indie*, while distinct in sound, share some overlap in features, which is reflected on the dendrogram.

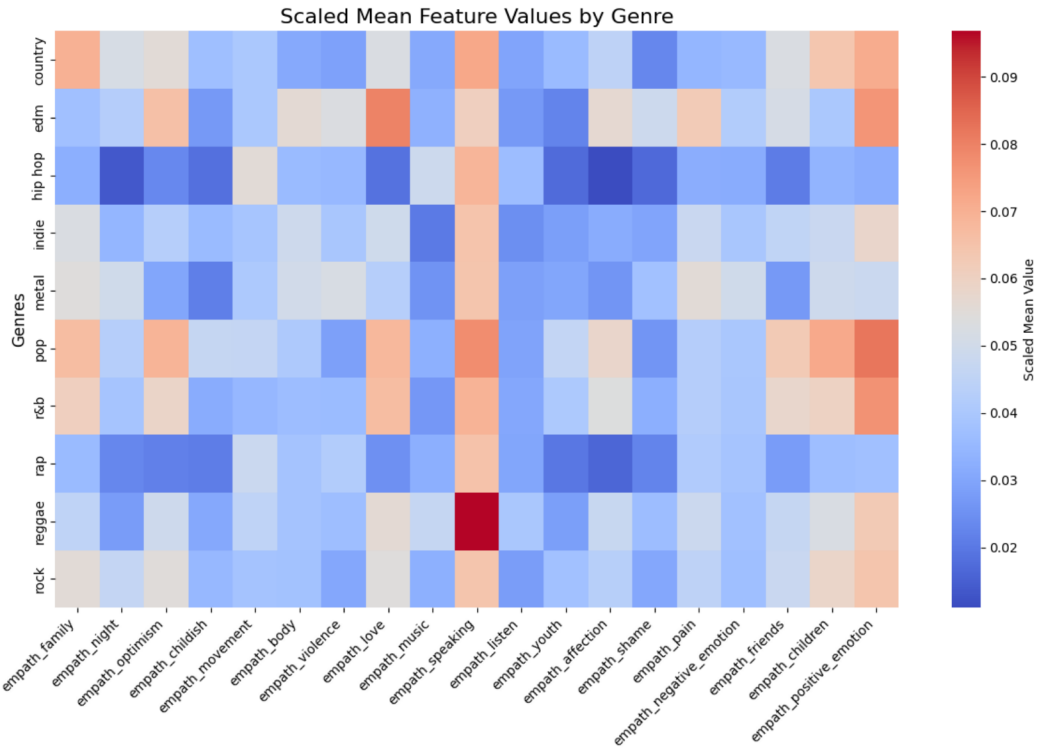

Figure 5.11: Heatmap of mean values of empath features in each genre

The heatmap displays scaled mean values of *Empath features* for different musical genres, highlighting thematic differences in their lyrics. It can be observed that:

- **Country** music lyrics often involve topics related to family, love and positive emotions.

It seems to align well with genre's tendency to narrate personal, heartfelt and often nostalgic stories.

- **EDM** lyrics seem to often touch upon love and optimism.
- For both **Hip Hop** and **Rap** according to Empath the most prominent topic is *speaking*.
- For **Metal** High values in *negative emotion* and *pain* underscore the genre's focus on intense, darker themes.
- Elevated values in *love* and *positive emotion* highlight **Pop** music's focus on themes of love and relationships.
- High values in *friends* and *optimism* categories reflect **Reggae's** emphasis on positivity, social connection, and uplifting messages.

5.5.2 Lyrical Similarity Based on Embeddings

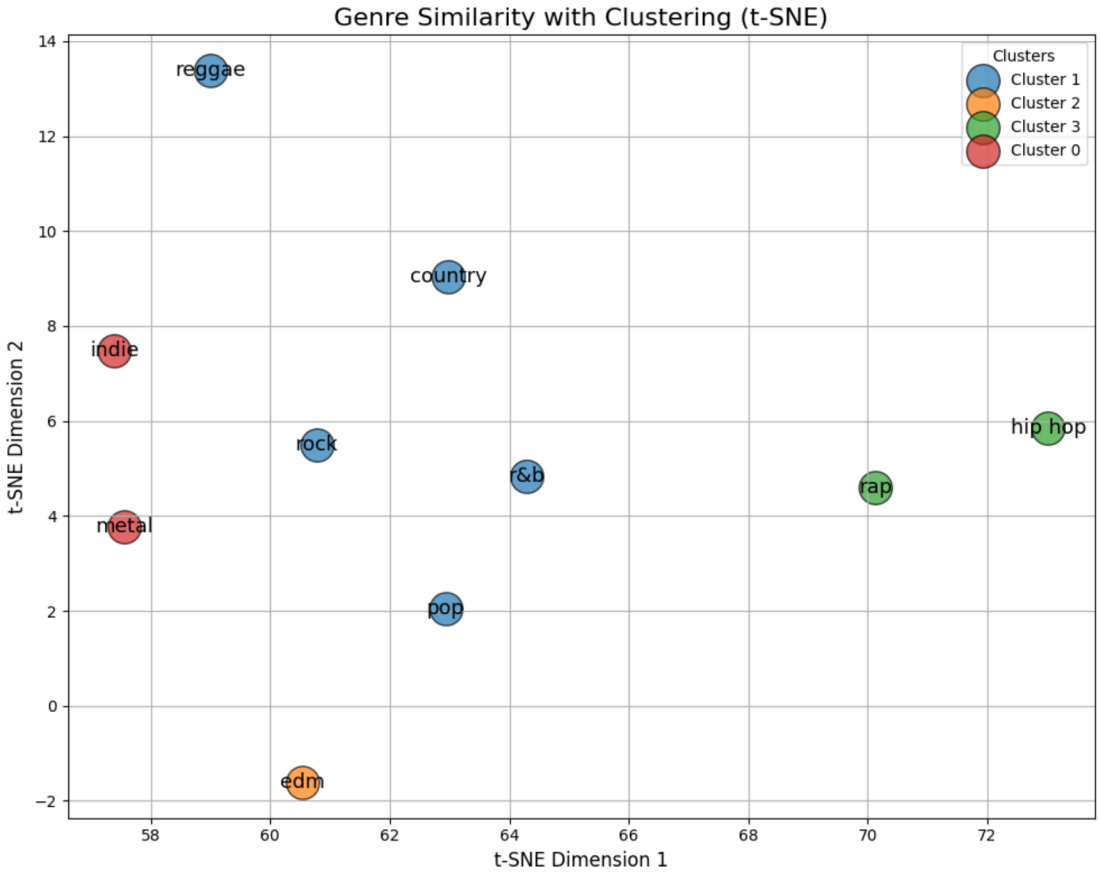

Figure 5.12: Heatmap highlighting the maximum feature values across genres.

t-SNE (t-Distributed Stochastic Neighbor Embedding) was applied to the Euclidean distance matrix of mean standardized Word2Vec and TF-IDF embeddings for each genre. This dimensionality reduction technique allows to project high-dimensional data into a 2D space while preserving the relative distances and similarities between data points as much as possible. This enables to visualize lyrical similarities between different genres.

In order to further explore lyrical relationships between the genres K-means clustering was applied on the mean standardized embeddings per genre,The optimal number of

clusters was determined using the elbow method, which suggested four distinct clusters. This additional operation clusters the most similar genres together.

- *Hip hop* and *rap* form a distinct cluster, indicating strong lyrical and thematic similarity.
- *Reggae* is slightly separate from other genres but still part of a larger cluster that includes *rock*, *pop*, *country*, and *R&B*, suggesting moderate similarity in lyrical and thematic content of songs belonging to those genres.
- *Metal* and *indie* are closely positioned, sharing overlapping themes while forming a separate cluster.
- *EDM* is positioned furthest from other genres, highlighting its unique lyrical and thematic style.

5.5.3 Top Genre Characteristics

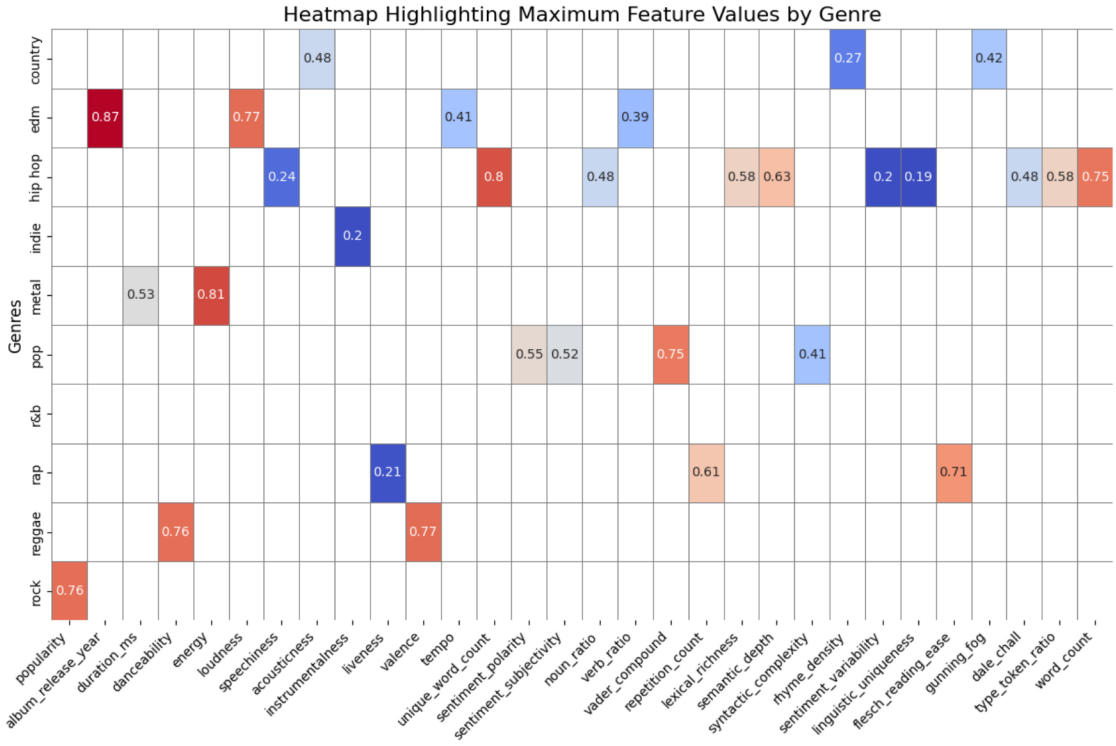

Figure 5.13: Heatmap highlighting the maximum feature values across genres.

Each filled cell represents the genre that exhibits the highest mean value for the corresponding feature, calculated using scaled feature values. This visualization emphasizes distinctive characteristics of each genre, and allows to find out "in which genre the feature achieved its highest mean".

- **Country**: Highest in acousticness and rhyme density
- **EDM**: Dominated in tempo, loudness, and featured the most recent songs
- **Hip Hop**: Excelled in unique word count, lexical richness, and semantic depth.
- **Metal**: Stood out for energy and longer song durations.

- **Pop**: Showed the highest positivity (VADER compound) and subjectivity in lyrics.
- **Rap**: Highest in repetition count and reading ease.
- **Reggae**: Highlighted by high valence and danceability.
- **Rock**: Scored highest in in popularity.

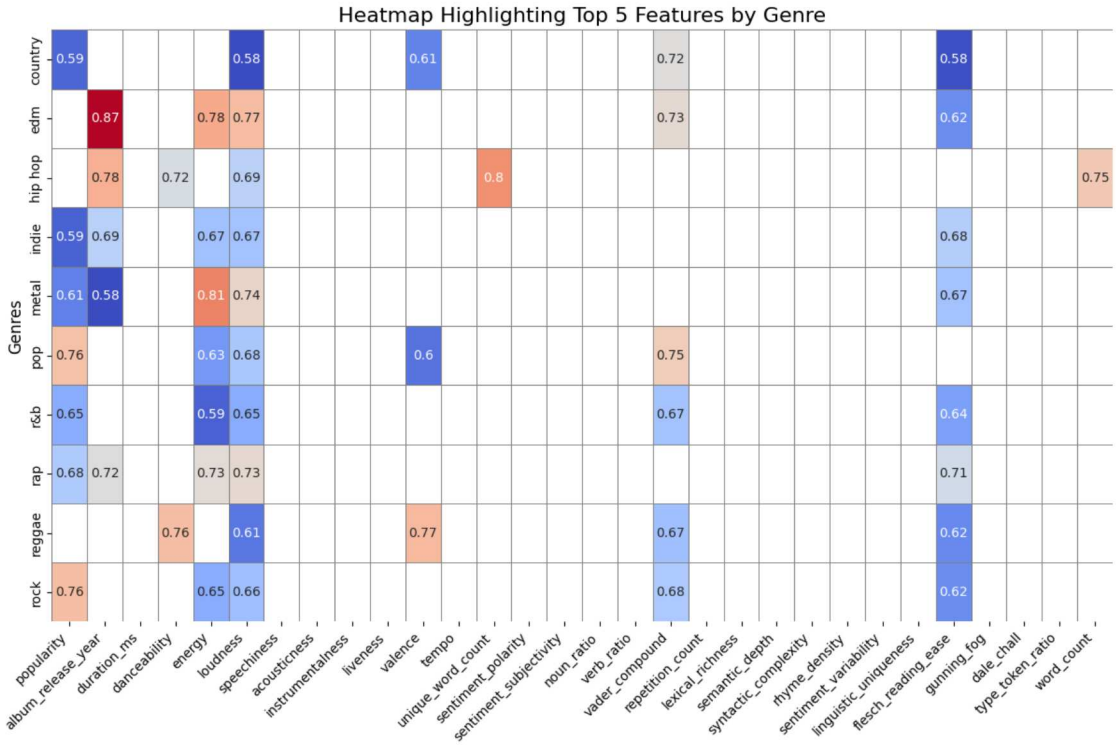

Figure 5.14: Heatmap highlighting the top 5 features by genre. Each filled cell represents one of the five highest mean feature values for a given genre, calculated on scaled data. This visualization focuses on identifying the most distinctive traits for each genre.

Lastly, we switch perspectives, focusing not on individual features but instead examining each genrer to identify the top 5 features that characterize it. Each filled cell in that heatmap shows one of the five most prominent characteristics exhibited by that genre.

6. Experiments and Results

6.1 Song Popularity

The analysis of song popularity provides valuable insights into the factors that influence the success of music tracks. This problem has two approaches:

- **Regression** - Spotify's popularity is a value on scale of 0-100. This approach involves training a regression model trying to predict that value.
- **Classification** - assigning binary label to the songs(popular vs. unpopular) and training a classification model to predict it.

In this section the prediction of popularity was attempted using several regression and classification models. These models were trained on different sets of features, including Spotify metadata, lyrical attributes, and audio features, with the aim of investigating the predictive power of those features and their impact on popularity.

Catboost models were used as the primary predictive tool due to their robustness and performance in handling complex relationships.

Baseline models were also implemented to serve as a reference point:

- *Baseline Mean Model*: baseline model for regression that always predicts the mean.
- *Baseline Majority Model*: baseline model for classification that always predicts the majority class.
- *Baseline Random Model*: baseline model for classification that predicts random class.

The experiments involved both quantitative evaluatoin and SHAP analysis to assess feature importance and interpretability of the models.

6.1.1 Regression Approach

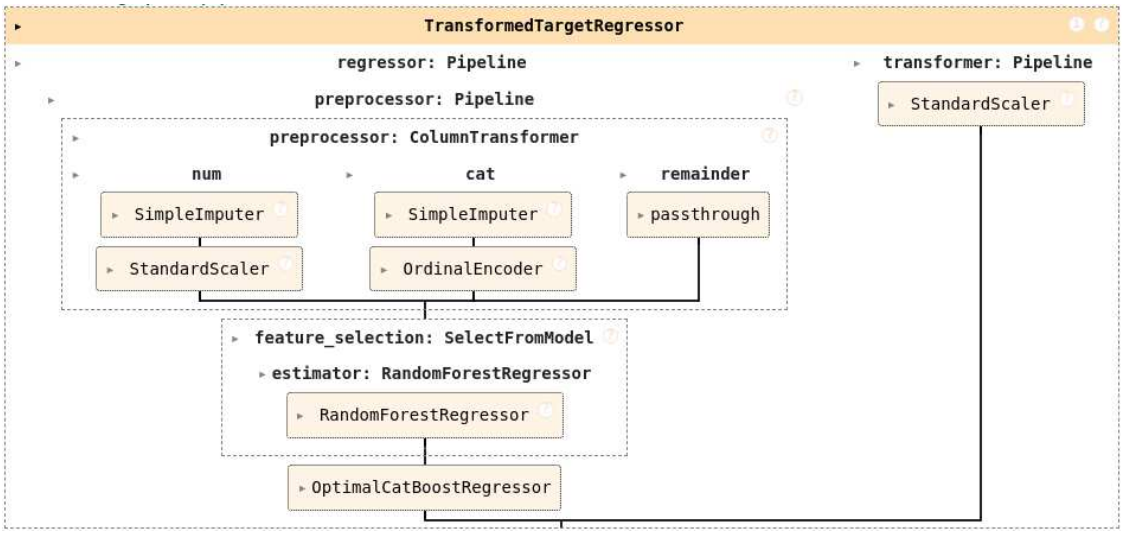

Figure 6.1: Regression model pipeline. It involves preprocessing, feature selection and CatBoost model.

The pipeline was fit with all available features, that includes Spotify audio features and metadata, lyrical features and features extracted from the audio files. The feature selection step chose the subset of most valuable features based on the feature importance from initial Random Forest model.

Tablica 6.1: Results of regression of popularity.

Model	**Features**	**MAE**	**RMSE**	R^2
Baseline Mean Model		13.53	16.90	-0.01
Catboost	all	**6.33**	**8.52**	**0.74**
Catboost	lyrical	11.81	15.29	0.17
Catboost	spotify data	**6.04**	**8.10**	**0.77**
Catboost	audio	12.71	16.15	0.08

As seen in the results table, the CatBoost model trained on spotify data only achieved the best performance, with **Mean Absolute Error(MAE) of 6.04 and and R^2 of 0.77**, closely followed by the model trained on all features that achieved **MAE of 6.33 and (R^2 of 0.74**. The results clearly indicate the importance of Spotify audio features and metadata in prediction of song's popularity. Predicting the success of a musical track based on solely the lyrics or acoustic features turned out to be diffifcult and models trained on those subsets of features showed only slight improvement in comparison to the baseline model. In contrast, the model trained on Spotify data only achieved **55.3%** better MAE score than the baseline model.

Despite feature selection, the model trained on all features performed slightly worse than the model trained on Spotify features alone. This likely occurred because adding less important or redundant features reduced the model's ability to focus on the most relevant Spotify features. The additional features may have introduced noise, making it harder to identify clear patterns. This behavior might also hint at slight overfitting on the redundant features, despite the use of cross-validation. Increased regularization could potentially mitigate this issue in future experiments.

SHAP analysis was conducted to interpret the predictions of the regression model and asses the importance of individual features and their contribution to song's popularity.

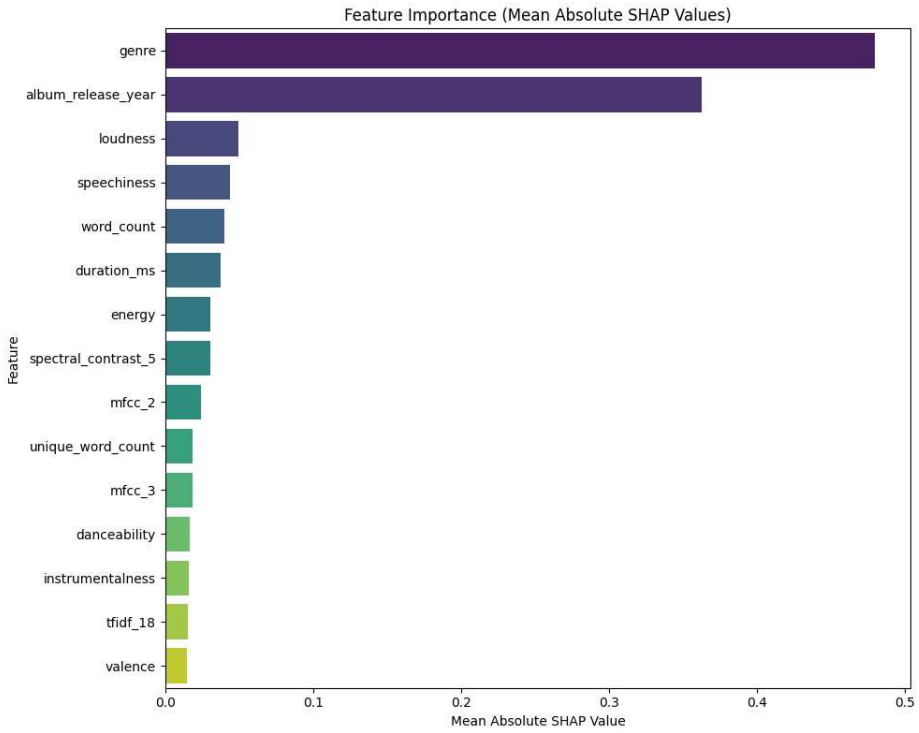

Figure 6.2: SHAP feature importance plot of the regression model for popularity trained on all features.

The mean absolute SHAP values were used to rank the overall importance of input features. Key insights:

- **Dominance of *Genre* and *Album Release Year.***: these two features account for the majority of the predictive power in the model. Genre captures the musical style and release year reflects trends and cultural preferences over time.
- Spotify's audio features like *loudness* and *speechiness* showed moderate contribution to model's performance. Their score highlights their relevance in describing popular tracks' audio characteristics.
- Lyric-based features like *word count* and *unique word count* were significantly less impactful, however still contributed to model's performance to some degree.

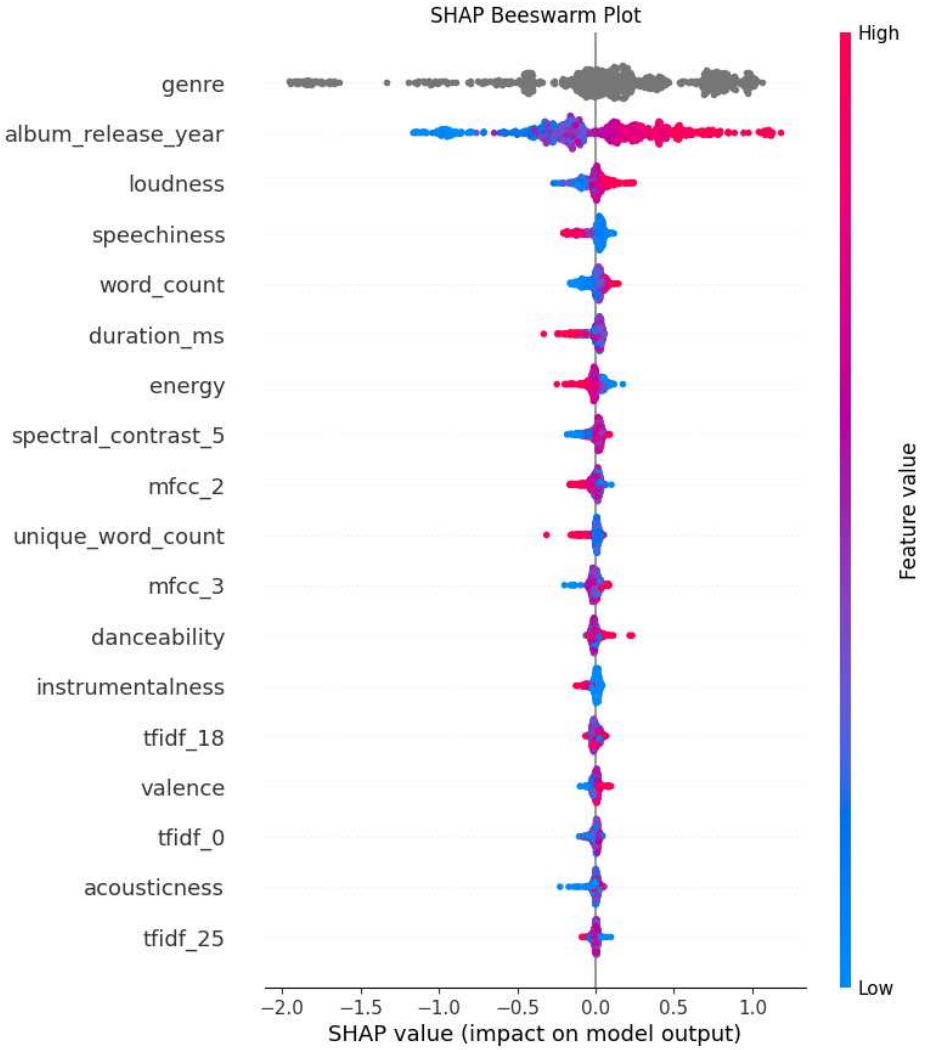

Figure 6.3: SHAP beeswarm plot of the regression model for popularity trained on all features.

- Higher values of *Release Year* strongly correlate with higher *Popularity*, indicating that newer songs are generally more popular.
- Higher *loudness* and lower *speechiness* were associated with increased *popularity*, reflecting their influence on listener preferences and music trends.

6.1.2 Classification Approach

The task of predicting song popularity was reformulated as a binary classification problem, where the target was to determine whether a song is "popular"(1) or ńot popular"(0). In order to create those labels from an integer variable with range 0-100, the **70th perrcentile of the popularity values was used as the threshold**. Songs with popularity greater or equal than this threshold werre labeled as popular, and the rest was labeled as unpopular. This thresholding approach based on quantile ensures a balanced representation of popular songs in the dataset while accounting for the naturally skewed distribution of popularity scores.

Tablica 6.2: Results of classification of popularity.

Model	**Features**	**Accuracy**	**F1(w.avg.)**
Baseline Majority Model		65.93%	52.39%
Baseline Random Model		48.27%	49.57%
Catboost	all	**84.68%**	**84.71%**
Catboost	lyrical	65.79%	64.77%
Catboost	spotify data	**82.06%**	**82.32%**
Catboost	audio	63.17%	63.00%

The classification models were evaluated using accuracy and weighted F1-score. *Baseline Majority Model* (which predicts the majority class for all samples) achieved an accuracy of 65.93% and an F1-score of 52.39%. This reflects the class imbalance introduced by the thresholding, with a higher proportion of songs being labeled as ńot popular."

The *Baseline Random Model*, which guesses classes randomly, performed worse in accuracy (48.27%) and achieved a slightly higher F1-score of 49.57% compared to accuracy due to its balanced attempts to handle both classes.

The CatBoost model significantly outperformed both baselines, achieving an overall accuracy of 84.68% and a weighted F1-score of 84.71% when trained on all features.This result demonstrates the effectiveness of the full feature set in capturing the complexity of

popularity prediction.

The CatBoost model trained on Spotify data performed well, with an accuracy of 82.06% and an F1-score of 82.32%, showing that Spotify features alone provide strong predictive power. However, the model trained on all features slightly surpassed it, suggesting that additional features offer some complementary information.

The model using only lyrical features achieved an accuracy of 65.79%, lower than the *Baseline Majority Model*, but a higher F1-score of 64.77%, reflecting its ability to handle class imbalance better than the majority model. This confirms that lyrical features are limited in their ability to predict song popularity.

The performance of the model trained on features extracted from audio files was similar to the lyrical one, demonstrating that isolated audio features are not sufficient to fully predict popularity.

SHAP analysis revealed that key features such as *genre*, *album release year* and *loudness* were the most significant predictors of popularity. These findings were consistent with the regression task, reinforcing the robustness of these features across different modeling approaches. This consistency highlights their central role in understanding and predicting song success.

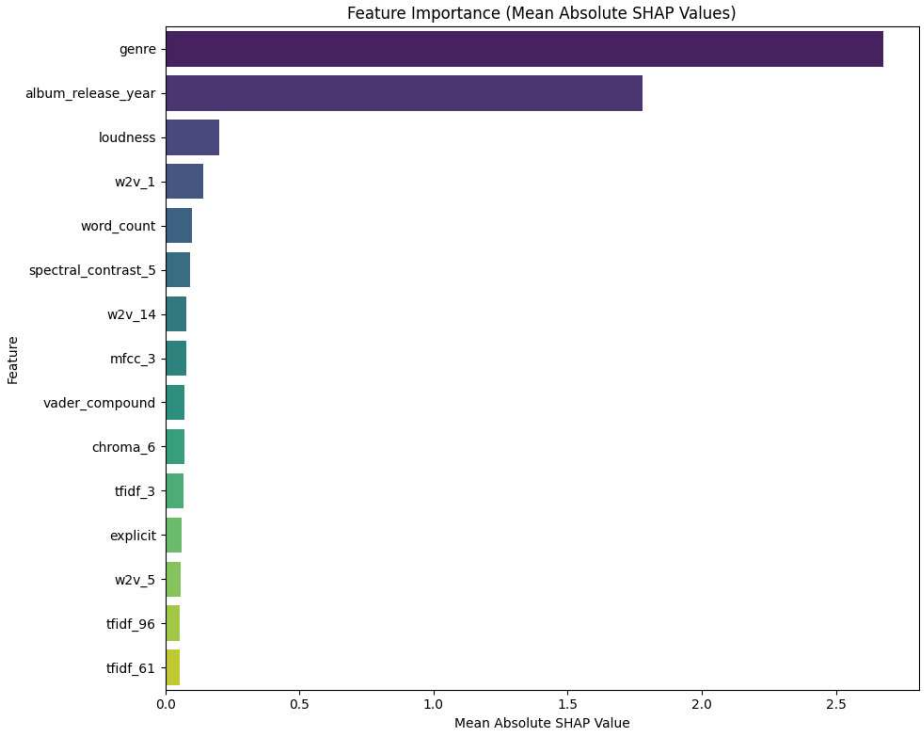

Figure 6.4: SHAP feature importance plot of the classification model for popularity trained on all features.

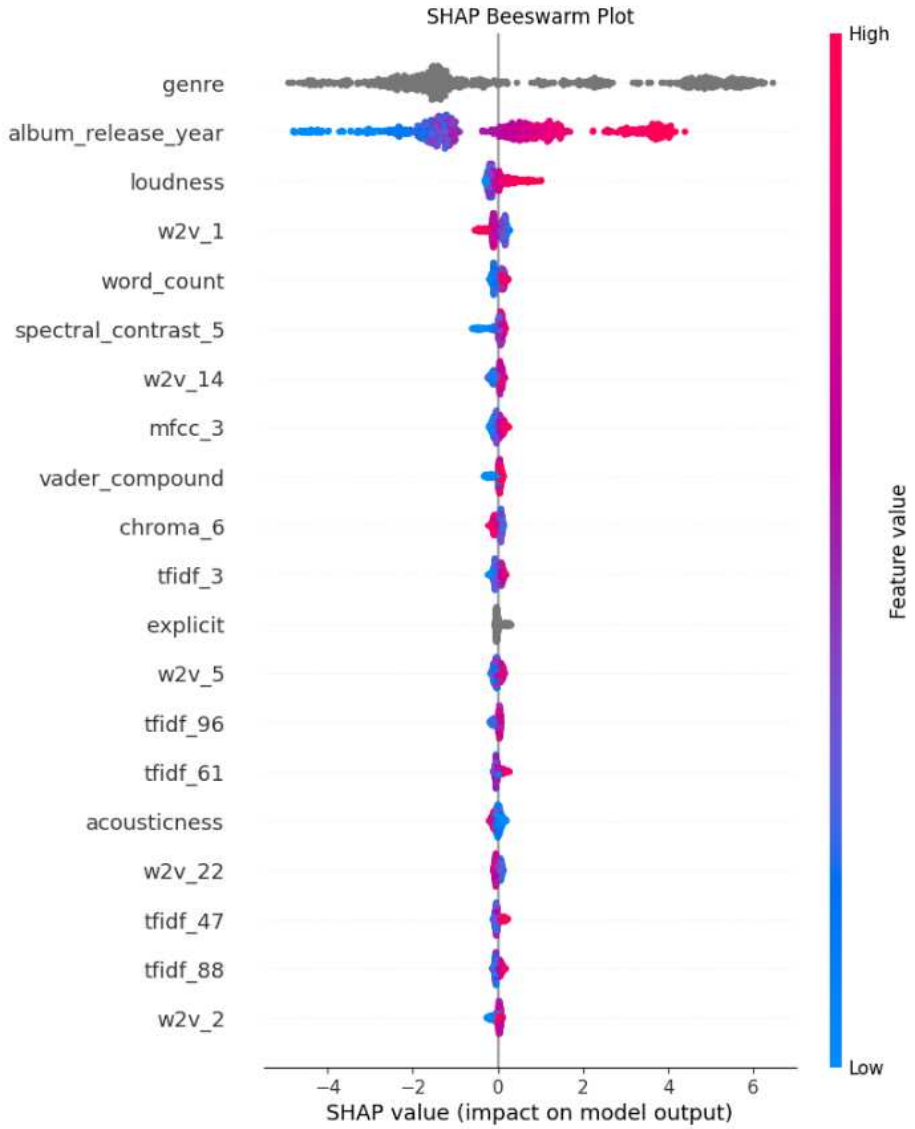

Figure 6.5: SHAP beeswarm plot of the classification model for popularity trained on all features.

The SHAP analysis for the classification model highlights similar dependencies to the regression model. Key features like *genre* and *album release year* remain the most important, emphasizing their role in shaping song popularity.

Several differences can be observed between feature importance in the regression and classification model, notably:

- Lyrics embeddings(TF-IDF and Word2Vec) seemed to play a bigger role in the classification model.
- Classification model seemed to pay much less attention to Spotify audio features, like *danceability*, *energy* and *speechiness*.
- Unlike in regression, in classification a sentiment metric, *VADER compound* contributed to the prediction of popularity. More popular songs tend to have more positive lyrics.

6.2 Explicitness

6.2.1 Classification Approach

The task of predicting whether a song contains explicit content was approached as a classification problem. The performance of the CatBoost model was compared against baseline on different feature subsets. One of the key challenges of this problem was significant class imbalance present in the dataset; approximately 85% of the songs did not contain explicit content.

Significant class imbalance can lead to models favoring the majority class, potentially resulting in high accuracy but poor performance on the minority class. To address this problem, CatBoost's class weights parameter was used. This parameter allowed the model to penalize misclassifications of the minority class more heavily, therefore improving its ability to recognize explicit content.

Tablica 6.3: Results of classification of explicitness.

Model	**Features**	**Accuracy**	**F1 weighted average**
Baseline Majority Model		84.00%	76.69%
Baseline Random Model		47.03%	53.91%
Catboost	all	**92.41%**	**92.40%**
Catboost	lyrical	**92.41%**	**92.25%**
Catboost	spotify data	86.48%	87.11%
Catboost	audio	82.75%	82.47%

The table presents the results of the models in terms of accurracy and weighted F1 score. The *Baseline Majority Model* achieved accuracy of 84% and significantly lower F1 weighted average of 76.69%, which reflects its inability to handle class balance effectively. The *Baseline Random Model* performed very poorly, with an accuracy of 47.03% and F1 weighted average of 53.91%.

In contrast, the CatBoost model outperformed both baselines achieving accuracy of **92.41%** and weighted average score of **92.4%** when trained on all features. Interestingly, the model trained only using lyrical features performed nearly as well, indicating that explicitness can largely be predicted based on lyrical content. Models that relied on Spotify metadata and features extracted from audio files had lower performance, achieving accurracy close to *Baseline Majority Model*, but with higher F1 scores, due to their ability to address class imbalance. These observations emphasize the centrality of lyrical information in prediction of explicit content.

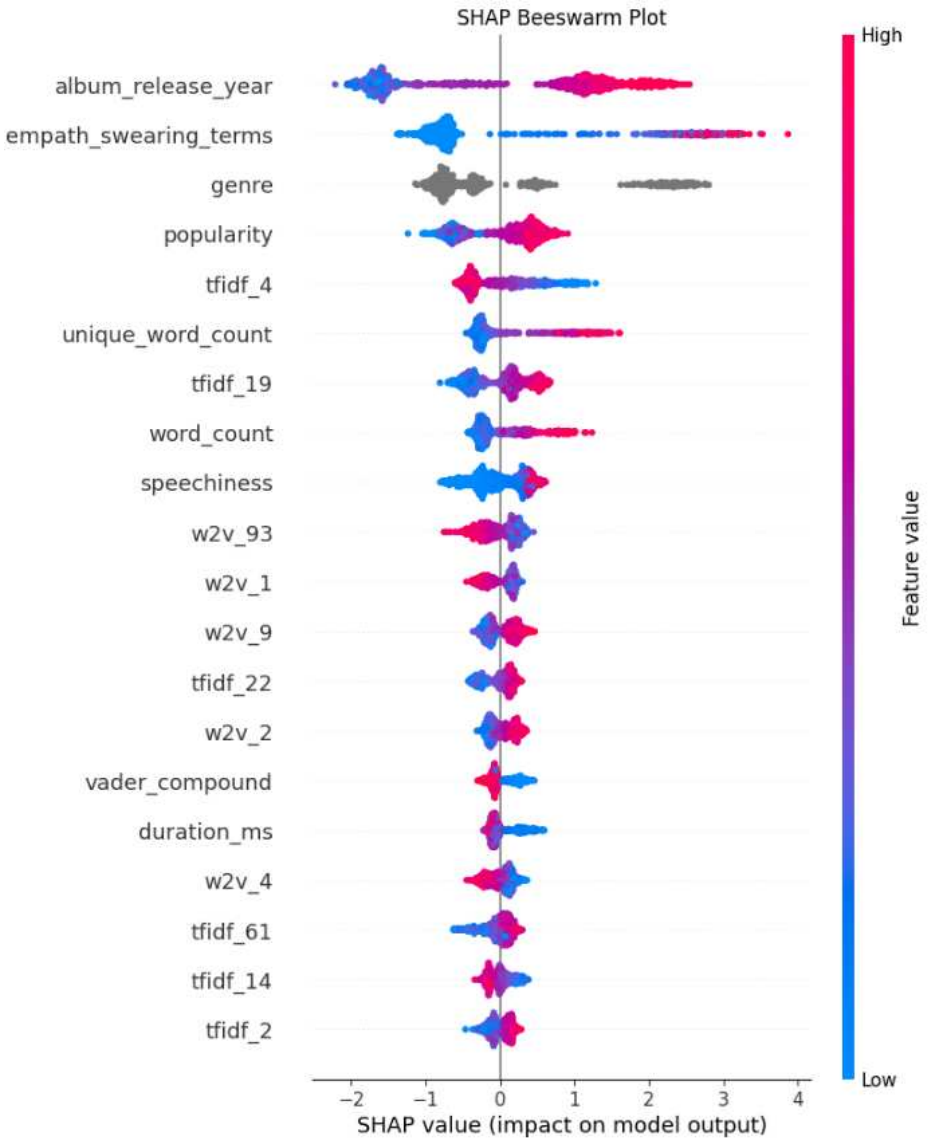

Figure 6.6: SHAP beeswarm plot of the classification model for explicitness trained on all features.

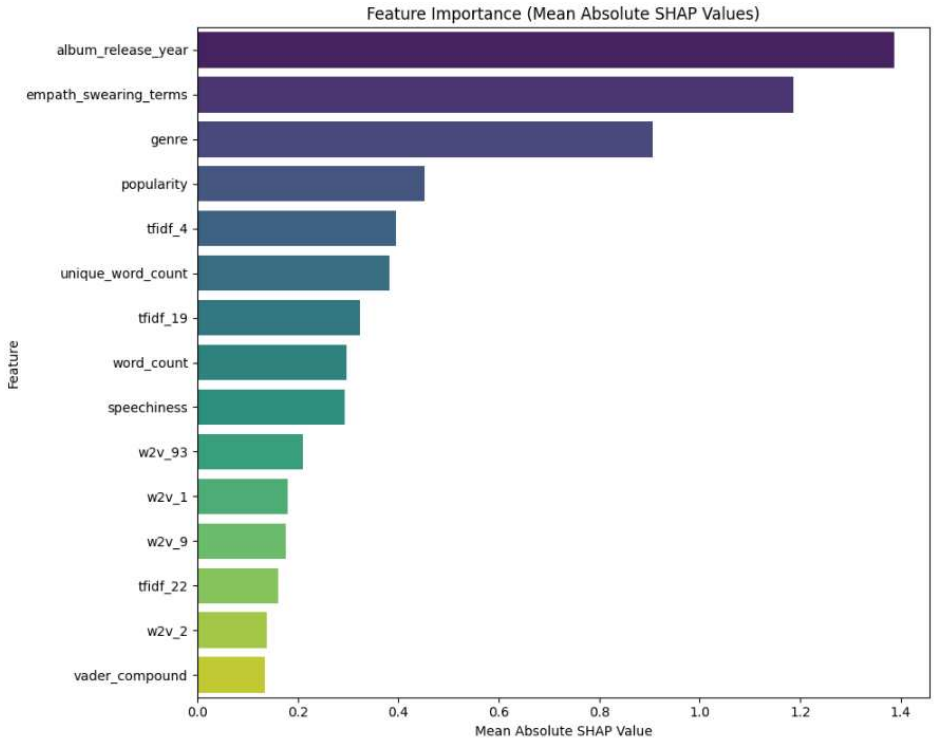

Figure 6.7: SHAP beeswarm plot of the classification model for explicitness trained on all features.

Based on the SHAP analysis we can observe that the key features for this task turned out to be:

- **Album Release Year**: this feature had highest feature importance, reflecting temporal trends in explicit content. Newer songs are statistically more likely to contain explicit language, reflecting changing societal norms and artistic expressions over time.
- **Empath Swearing Terms**: As expected, the presence of swearing terms was a strong indicator of explicitness.
- **Genre** was the third most important feature, showing that certain genres are more likely to contain explicit content than others.
- **TF-IDF**: several vectors created using TF-IDF and later reduced via PCA significantly

contributed to the model's predictions. Similar to the Empath feature, TF-IDF likely captured swear words and related patterns in the lyrics, reinforcing its importance.

- **Speechiness**: songs with higher values of speechiness—indicating a greater presence of spoken-word elements—were more likely to be labeled as explicit.

The task of predicting explicitness could potentially achieve higher performance by leveraging the interpretability of explicitness, which provides clear hints about the most relevant features for this property. However, the primary goal of this analysis was not to maximize model performance but to explore the impact of various feature subsets on the prediction of explicitness.

The results indicate that predicting explicitness based solely on audio features is inherently challenging. The model trained on audio features performed worse than the baseline, suggesting that the extracted audio characteristics from the MP3 files do not sufficiently capture the explicit nature of songs. This highlights the need for features more directly related to lyrical or contextual information when modeling explicit content.

6.2.2 Impact of Explicit Language on Popularity and Sentiment

The relationship between explicit content and song popularity is a topic of interest in understanding the cultural and commercial dynamics of music. Explicit songs often reflect bold themes, which may resonate more strongly with certain audiences.

To investigate this, a bootstrap test was conducted to determine whether explicit songs are, on average, more popular than non-explicit songs. Bootstrap was chosen for this analysis due to its robustness and flexibility. Unlike traditional parametric tests, the bootstrap method does not rely on strong assumptions about the underlying data distribution, making it well-suited for analyzing real-world datasets that may not meet normality or other strict requirements.

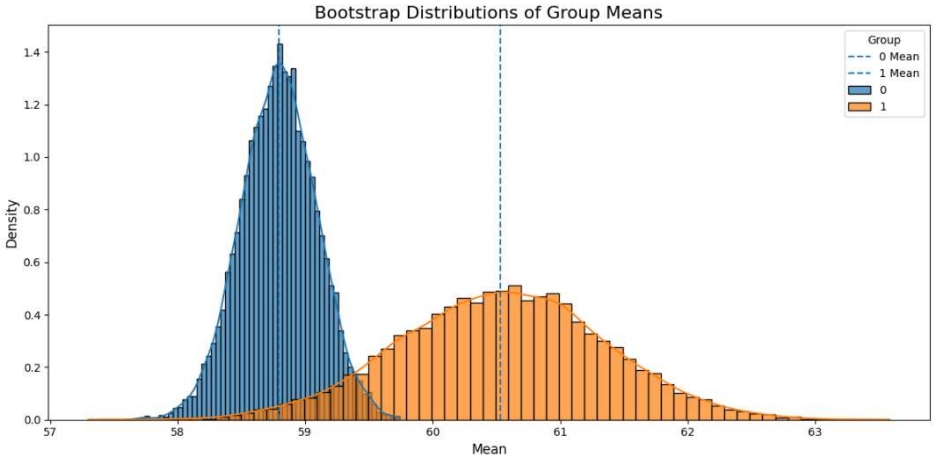

Figure 6.8: Bootstrap Test to check if explicit songs are on average more popular.

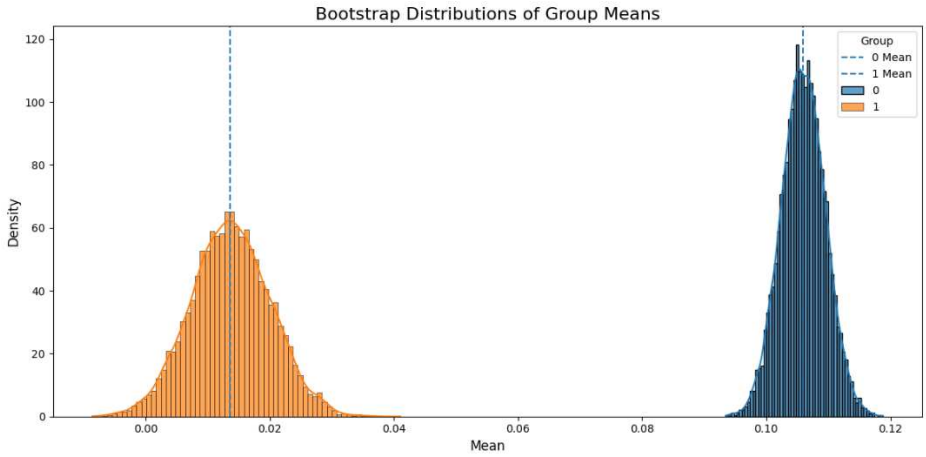

Figure 6.9: Bootstrap Test to check if explicit songs are on average more positive or negative than non-explicit songs.

Tablica 6.4: Results of the Bootstrap Test.

Metric	**Control Group**	**Treatment Group**	**pct_diff**	**pct_ci_lower**	**pct_ci_upper**	**abs_diff**	**abs_diff_ci_lower**	**abs_diff_ci_upper**
Popularity	non-explicit	explicit	**2.96%**	0.08%	5.79%	1.74	0.0453	3.4013
Sentiment Polarity	non-explicit	explicit	**-87.10%**	-100.69%	-73.63%	-0.09	-0.1067	-0.0780

The analysis reveals that explicit songs are, on average, **2.96% more popular** than non-explicit songs. The confidence interval of [0.08%, 5.79%] suggests statistically significant but modest increase in popularity for explicit songs.

For sentiment polarity, explicit songs turned out to be on average **87.10% less positive** than non-explicit songs. The confidence interval of [-100.69%, -73.63%] suggests a very significant difference. This observation aligns with the notion that bold songs that contain explicit language often reflect intense or provocative themes, which may carry less positive emotion.

6.3 Sentiment

Figure 6.10: Wordclouds of Positive and Negative Sentiment Songs.

This section tackles the problem of song lyrics sentiment prediction. The binary sentiment labels(positive (1) and negative (0)) were derived from the variable *sentiment polarity*.

Songs with polarity values below 0 were labeled as negative, while those above 0 were labeled as positive.

This process resulted in an unbalanced target variable, where the positive class was overrepresented. Additionally, due to the left-skewed distribution of *sentiment polarity*, songs labeled as *positive* were, on average, more strongly positive compared to the relatively moderate negativity of songs in the *negative* class.

While this imbalance and class definition might pose challenges for traditional predictive modeling, it is not a major concern for this study. The primary objective is to understand which factors contribute to a song's sentiment rather than achieving perfect predictive accuracy. Chosen approach provides sufficient insight into the relationship between song features and sentiment.

Tablica 6.5: Results of classification of sentiment.

Model	**Features**	**Accuracy**	**F1(w.avg.)**
Baseline Majority Model		69.65%	57.19%
Baseline Random Model		52.55%	54.46%
Catboost	all	71.44%	69.77%
Catboost	lyrical	**71.86%**	**70.72%**
Catboost	spotify data	67.44%	64.94%
Catboost	audio	65.10%	63.43%

The *Baseline Majority Model*, which simply predicted the most common class for all samples, achieved an accuracy of 69.65% and an F1-score of 57.19%, reflecting the class imbalance in the dataset. The *Baseline Random Model*, which predicted classes randomly, performed worse in terms of accuracy (52.55%) but slightly better in F1-score (54.46%).

Among the CatBoost models, the one trained solely on lyrical features performed the best, achieving an accuracy of 71.86% and a weighted F1-score of 70.72%. This highlights the importance of lyrical features in capturing sentiment. The model trained on all features followed closely, but performed slightly worse due to large amount of redundant features, despite the presence of feature selection step in the pipeline that attempted to discard them.

Models trained on only Spotify or audio features performed worse, achieving accuracies of 67.44% and 65.10%, respectively. These results suggest that while Spotify and audio features capture useful information, they are less indicative of sentiment compared to lyrical data.

The results indicate that features extracted directly from MP3 files are not strong predictors of a song's sentiment, despite the intuitive assumption that audio characteristics should play a role. This outcome suggests that the current set of extracted audio features may not sufficiently capture the nuanced aspects of sentiment. Future work could focus on advanced feature engineering or exploring additional audio-based descriptors to improve the performance of audio-based sentiment prediction models.

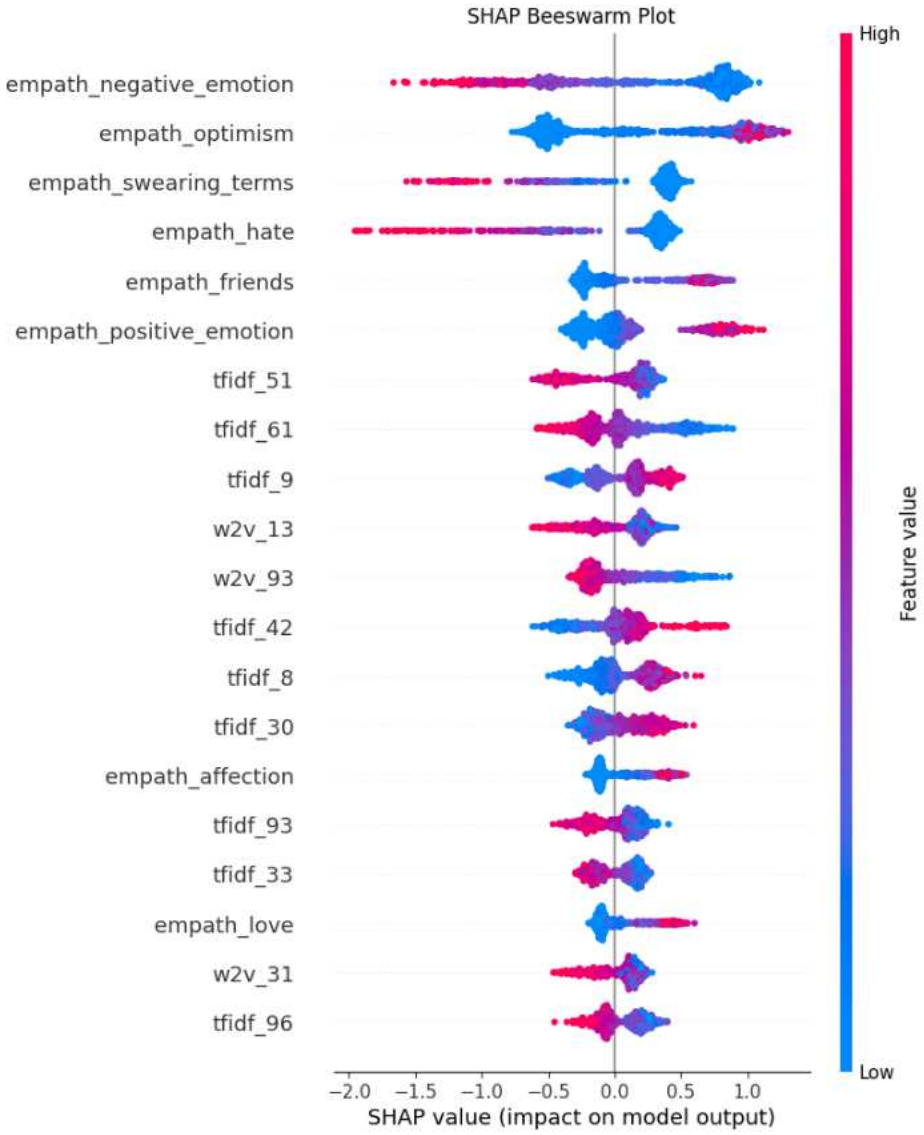

Figure 6.11: SHAP beeswarm plot of the classification model for sentiment trained on all features.

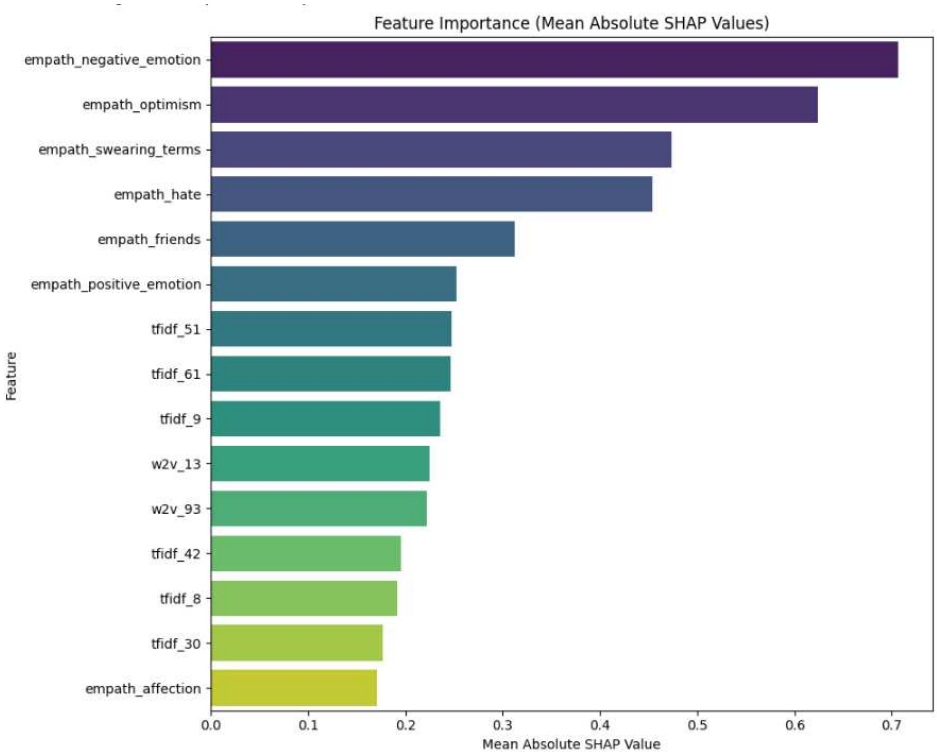

Figure 6.12: SHAP feature importance plot of the classification model for sentiment trained on all features.

The SHAP analysis showed that lyrical features, like *empath negative emotion*, *empath optimism*, and *empath swearing terms*, were the most influential in predicting sentiment, with negative emotions and swearing linked to negative sentiment, and optimism and positive emotion tied to positive sentiment. TF-IDF and Word2Vec embeddings also contributed, reflecting patterns in lyrics. Audio features, however, had minimal impact, emphasizing the need for improved audio feature engineering to capture sentiment.

6.4 Genre

This task aimed to predict song's genre using various feature subsets. The dataset included 10 distinct genres, covering a diverse range of musical styles.

Tablica 6.6: Results of classification of genre.

Model	**Features**	**Accuracy**	**F1(w.avg.)**
Baseline Majority Model		11.17%	02.24%
Baseline Random Model		10.20%	10.38%
Catboost	all	**61.24%**	**60.27%**
Catboost	lyrical	32.96%	32.23%
Catboost	spotify data	56.96%	56.65%
Catboost	audio	31.31%	30.33%

The baseline models achieved very poor performance. The *Baseline Majority Model* achieved accuracy of 11.17% and weighted F1 of 2.24%, and the *Baseline Random Model* accuracy of 10.20% and F1 of 10.38%.The CatBoost model trained on all features significantly outperformed the baselines, achieving 61.24% accuracy and 60.27% weighted F1-score. This highlights the importance of combining diverse feature sets (lyrics, Spotify data, audio features, embeddings) for genre prediction.

Among the individual feature subsets, Spotify features turned out to be the most predictive, resulting in a relatively well-performing model. While lyrical and audio features alone were not strong predictors of genre, their combination with Spotify features significantly improved the model's performance, highlighting the complementary nature of these data sources.

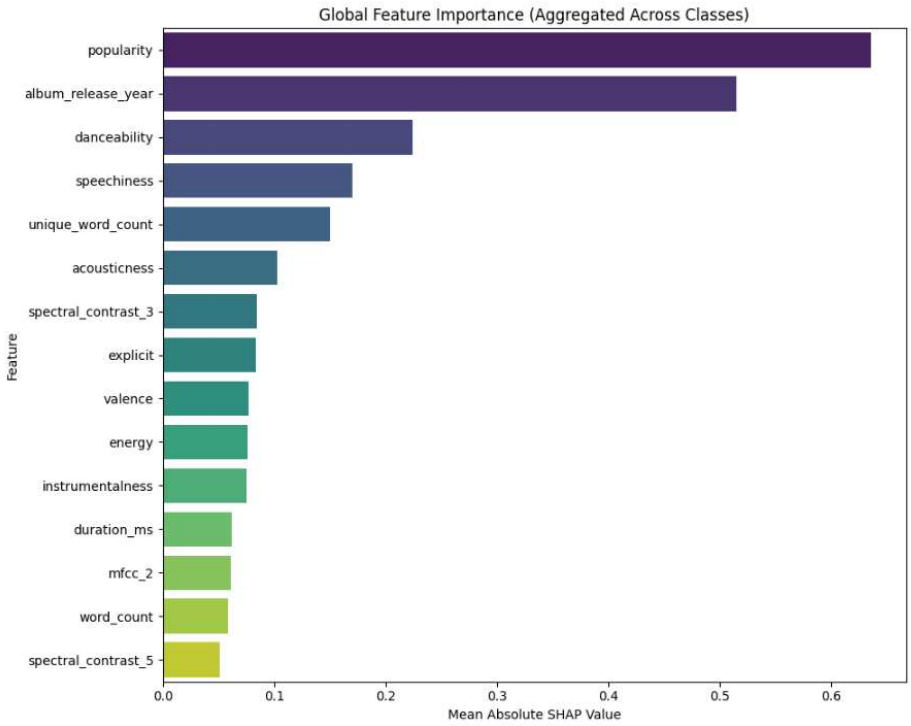

Figure 6.13: SHAP feature importance plot of the classification model for genre trained on all features. The scores were averaged across all classes.

Looking at the SHAP global feature importance plot(values averaged across all classes) we can observe that:

- *Popularity* and *album release year* emerged as the most critical predictors of *genre*. These two features likely capture trends and listener preferences linked to specific genres.
- Other key features included *danceability*, *speechiness*, and *unique word count*, which relate to the style of the music and the content of the lyrics, helping to distinguish between genres.
- Audio features like *acousticness* and *spectral contrast* also played a significant role, highlighting that the sound of the music reflects its genre.

- The *explicit* flag also had a noticeable effect, which makes sense given its strong connection to genres like rap and hip-hop.

6.5 Topics Modelling

6.5.1 Latent Dirichlet Allocation

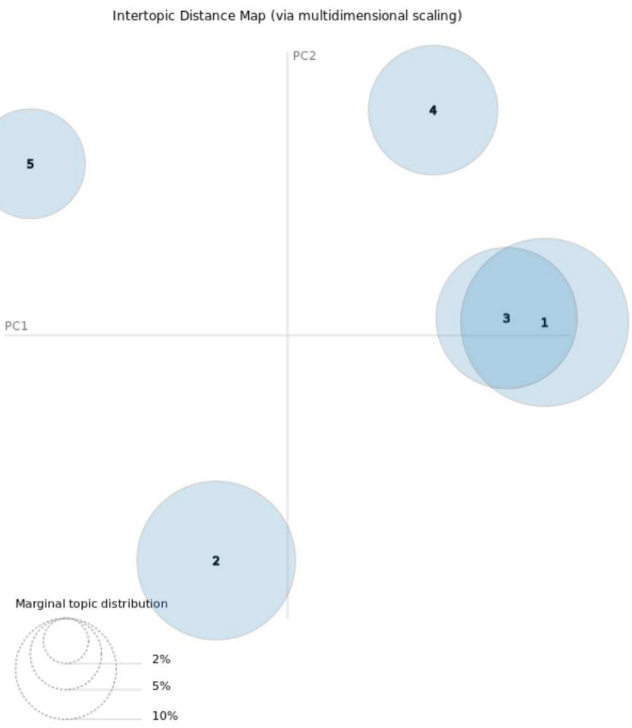

Figure 6.14: Intertopic Distance Map.

- The bubble chart shows the intertopic distance map using multidimensional scaling (MDS). Each bubble represents a topic extracted by the LDA model.
- The position of the bubbles reflects the relationship between topics. Topics closer together share more words or themes in common.

- The size of the bubbles represents the prevalence of the topic within the entire dataset. Larger bubbles correspond to topics with a higher proportion of terms in the dataset.

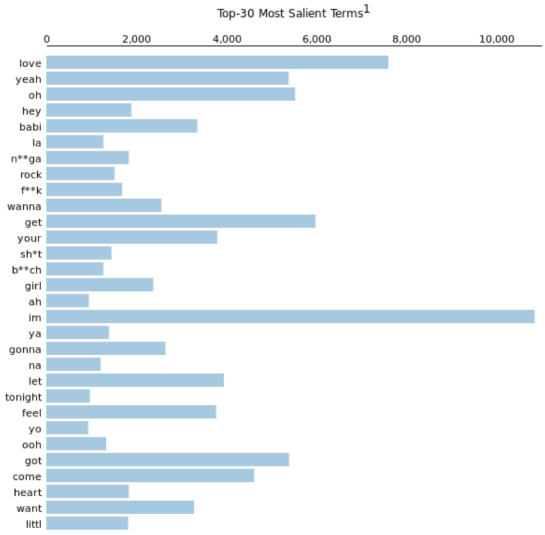

Figure 6.15: Top-30 Most Salient Terms across the entire corpus. The bars represent how often each term appears in the entire corpus.

Blue bars show overall frequency of the terms in the entire corpus. Red bars indicate estimated frequency of the terms in that topic. Now we can look at the results in each topic:

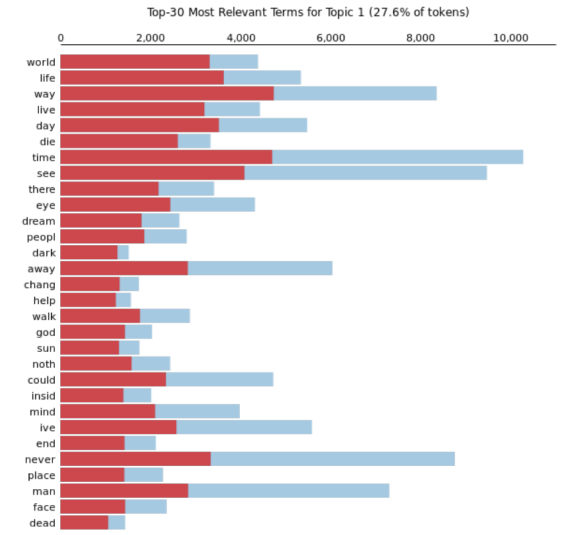

Figure 6.16: Top-30 Most Relevant Terms in Topic 1.

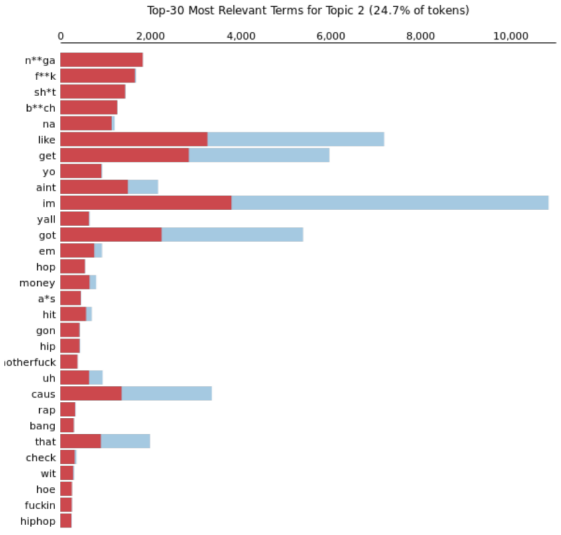

Figure 6.17: Top-30 Most Relevant Terms in Topic 2.

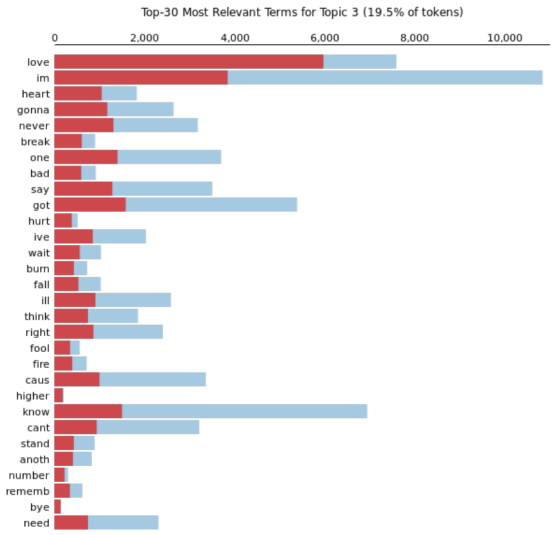

Figure 6.18: Top-30 Most Relevant Terms in Topic 3.

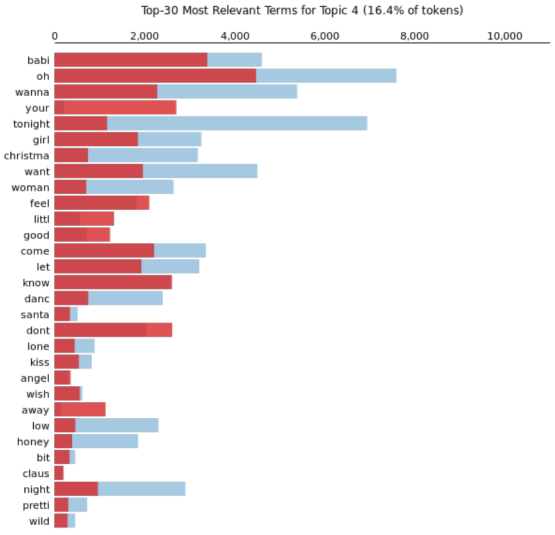

Figure 6.19: Top-30 Most Relevant Terms in Topic 4.

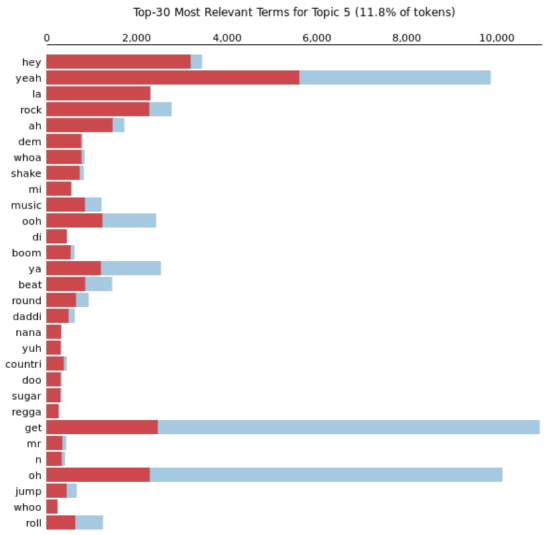

Figure 6.20: Top-30 Most Relevant Terms in Topic 5.

The topic modeling results reveal distinct themes found in the analyzed song lyrics. The intertopic distance map(Figure 6.14) shows that the five topics are well separated(except for 1 and 3, that have a significant overlap). The model has successfully captured distinct clusters of related terms, providing valuable insight and observations:

- Topic 1 appears to revolve around **existential and philosophical themes**, with frequent use of words like "world", "life", "time", "die", "live", and "dream".
- Topic 2 is characterized by **colloquial and explicit language**, exhibiting themes related to **rap and hip hop culture**. This topic is strongly tied to urban and street culture themes.
- Topic 3 revolves around themes of **love, heartbreak, emotional conflict and relationships**, as indicated by words like "love", "heart", "hurt, "break".
- Topic 4 focuses on themes of **romantic affection, celebration, and festives**. Words such as "baby", "tonight", "girl", "wanna", and "kiss" highlight intimate emotions and the inclusion of "Christmas", "Santa" and other related words strongly suggests a Christmas-themed music.

- Topic 5 appears to center around themes of **music, rhythm and energetic expressions**. Most prominent words identified for that topic are “hey“, “yeah“, “rock“, “music“ and “beat“. They suggest a focus on the sound and feel of music. Inclusion of terms such as “regga“(stemmed version of “reggae“) indicates possible inclusion of reggae songs.

The topic modeling analysis effectively uncovered five distinct thematic clusters within the song lyrics dataset, each representing unique aspects of lyrical expression, ranging from love and affection, to urban culture, and Christmas themes. The intertopic distance map (Figure 6.14) confirms that the identified topics are well-separated, with overlaps only in related themes such as existentialism (Topic 1) and emotional conflict (Topic 3).

6.5.2 Genre Distribution Across Topics

These results demonstrate the potential of topic modeling to extract and analyze dominant lyrical themes in music at scale. After assigning the most probable topic to each song the distribution of features and genres in each topic can be checked.

Figure 6.21: Heatmap showing percentage of genres in each topic.

The heatmap provides a detailed representation of the percentage distribution of genres within each topic derived from the LDA model. Key observations include:

- **Topic 1** has a balanced distribution across multiple genres, with slight overhead of *metal*, that makes up for over 20% of the songs in that topic.
- **Topic 2** is dominated by *rap* and *hip hop* genres, which aligns with its themes of urban culture and colloquial language.
- **Topic 3** focuses on love and heartbreak, a common theme in music across many genres. The heatmap doesn't show any dominant genre, with *R&B*, *rock*, *pop*, and *country* making the highest contributions.
- **Topic 4** has a significant representation in *R&B*, *pop* and *country*, suggesting its romantic and festive focus, including elements like celebrations and intimate moments.
- **Topic 5**: strongly linked to *reggae* and *rock*, reflecting the energetic and rhythmic themes of music and sound.

Based on the heatmap topics 1 and 3 show more diverse genre distribution, indicating thematic universality or overlap. This observation aligns with the conclusions drawn from the distance map(Figure 6.14) and the analysis of their most relevant terms((Figure 6.16, (Figure 6.18)). Topics 2 and 5 display strong dominance by specific genres, showing more focused thematic expressions.

6.5.3 Empath Features Distributions Across Topics

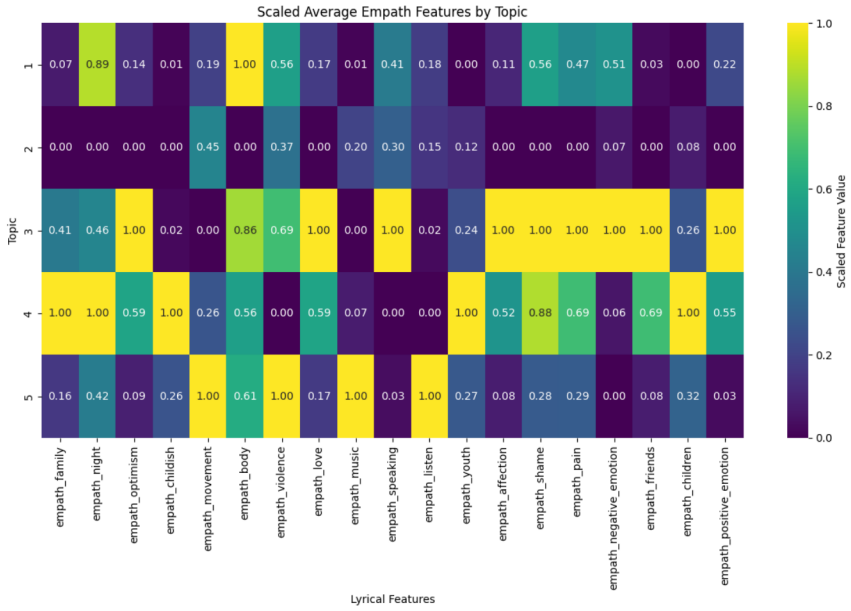

Figure 6.22: Heatmap showing the distribution of scaled empath features across topics.

This heatmap visualizes the scaled average Empath features by topic, highlighting the significanace of specific lyrical themes within each topic. Key observations include:

- **Topic 1**: high values of *night*, *body*, *pain* and *shame* suggest lyrics focusing on darker or more aggresive themes, aligning well with the dominance of metal as the main genre.
- **Topic 2**: strong association with *violence*, *movement*, *speaking* and *youth*, which are themes frequently occurring themes in rap and hip hop music.
- **Topic 3**: Strong emphasis on *love*, *affection*, *negative emotions*, and *friends* reflects themes of relationships, heartbreak, and emotional conflict.
- **Topic 4**: High values for *family*, *children*, *positive emotions*, and *youth* suggest themes of affection and relationships with loved ones, possibly relating to Christmas or other festive occasions.

- **Topic 5**: High values in *music*, *movement*, and *violence* suggest rhythmic and energetic themes.

6.5.4 Sentiment Across Topics

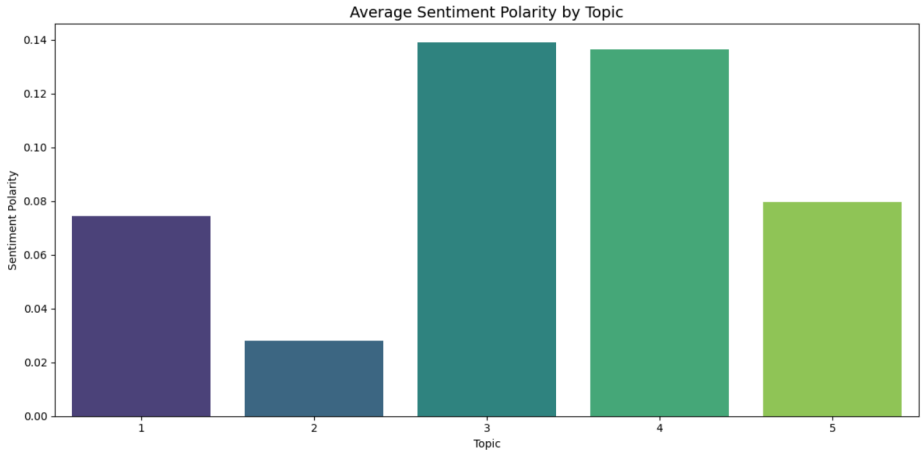

Figure 6.23: Bar chart showing average sentiment polarity by topic.

The bar chart illustrates the average sentiment polarity for each topic. Topics 3 and 4 exhibit the highest sentiment polarity, indicating more positive themes, while Topic 2 has the lowest polarity, reflecting more neutral or potentially negative sentiment, which intuitively aligns with previous findings.

6.5.5 Spotify Features Across Topics

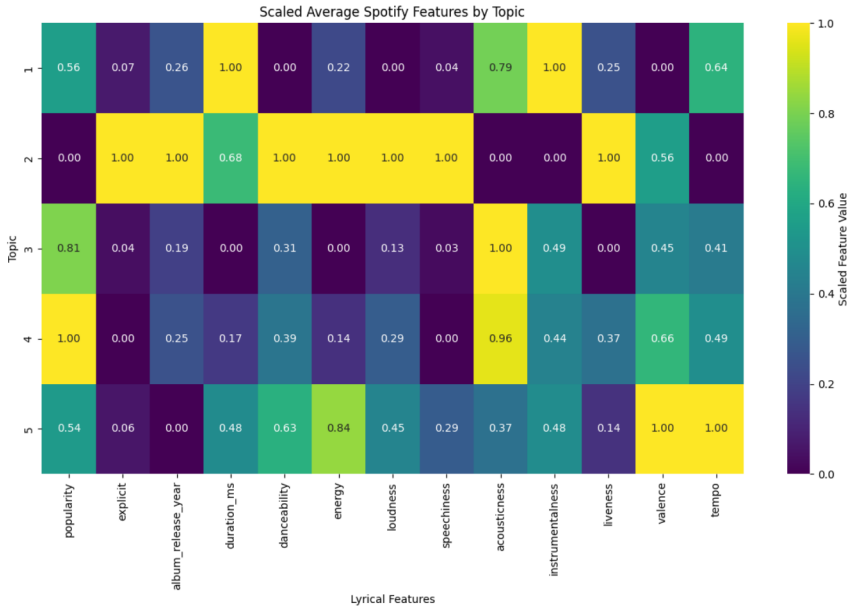

Figure 6.24: Heatmap showing average values of scaled spotify features in each topic.

The heatmap provides a visualization of the average scaled Spotify features across topics. It shows distinct patterns for each topic, highlighting their unique musical attributes:

- **Topic 1**: High values in *acousticness* and *instrumentalness*, paired with low *danceability* and *valance* suggest calmer, more instrumental tracks.
- **Topic 2**: High scores in *explicitness*, *speechiness*, *liveness*, *energy* and *loudness* indicate tracks with s trong vocal content, high energy and explicit language, aligning well with the dominant genres in this topic.
- **Topic 3**: High *popularity* and *acousticness*, paired with low *energy* and *danceability* suggest slower, more emotional, acoustic tracks.
- **Topic 4**: High *acousticness*, *popularity*, relatively high *valence* and *tempo* indicate s ofter, somewhat upbeat tracks that include both instrumental and vocal components.

- **Topic 5**: High values in *tempo*, *valence*, and *danceability*, along with moderate *energy*, suggest lively, upbeat, and energetic and possibly danceable tracks.

6.6 Temporal Trends in Music

To analyze the temporal trends in musical features, during the data collection phase the variable *album release year* was stratified into 10-year buckets representing distinct decades. This stratification allowed for a comparative analysis of feature evolution across time periods.

6.6.1 Identification of Features Affected by Release Year

The process for identifying features influenced by release year involved two main steps:

- **Linear Regression Analysis**: Simple linear regression models were trained for each feature using the decade as the independent variable. The coefficient of determination (R^2) values were extracted to quantify how well the decade explained the variance in each feature.
- **ANOVA Testing**: Analysis of variance (ANOVA) was conducted to assess the statistical significance of differences in feature means across decades. The resulting p-values were used to identify features where temporal variation was statistically significant.

Features exhibiting high R^2 values and statistically significant p-values ($\alpha = 0.05$) were selected as candidates for further analysis.

24 features were identifed to have meaningful dependency on the decade of release, those are their plots along with 95% confidence intervals:

A total of 24 features were identified as having a meaningful dependency on the decade of release. These features demonstrated statistically significant p-values and relatively high R^2 coefficients, indicating notable changes in their values over time.

Tablica 6.7: Features identified as having significant dependency on the decade of release.

Feature	**R^2**	**Trend Coefficient**	**F-Statistic**	**p-Value**
explicit	0.909073	0.006645	76.023316	0.000000
valence	0.906611	-0.003028	45.487845	0.000000
syntactic_complexity	0.905622	0.519760	3.188473	0.002285
popularity	0.893476	0.276672	61.320459	0.000000
vader_compound	0.880113	-0.007787	11.830054	0.000000
sentiment_polarity	0.853502	-0.001479	8.135521	0.000000
loudness	0.833296	0.068308	139.657491	0.000000
rhyme_density	0.824603	-0.000094	18.454532	0.000000
danceability	0.814174	0.001072	7.099057	0.000000
speechiness	0.806721	0.000817	18.774707	0.000000
dale_chall	0.775952	0.031306	5.824913	0.000001
flesch_reading_ease	0.771849	-0.415172	3.364072	0.001401
gunning_fog	0.758302	0.154463	3.178465	0.002350
lexical_richness	0.565358	-0.000709	10.507191	0.000000
type_token_ratio	0.565358	-0.000709	10.507191	0.000000
energy	0.518081	0.003036	84.094129	0.000000
acousticness	0.469885	-0.005867	221.629618	0.000000
duration_ms	0.151967	495.225838	102.202431	0.000000
semantic_depth	0.115761	0.000821	3.334177	0.001523
tempo	0.110478	0.036494	2.943202	0.004482
sentiment_variability	0.064628	-0.000182	5.719228	0.000001
linguistic_uniqueness	0.020538	0.000049	2.558628	0.012530
instrumentalness	0.016573	0.000118	4.748720	0.000025

The following plots display the trends of these features across decades, along with 95% confidence intervals to illustrate the variability of the trends:

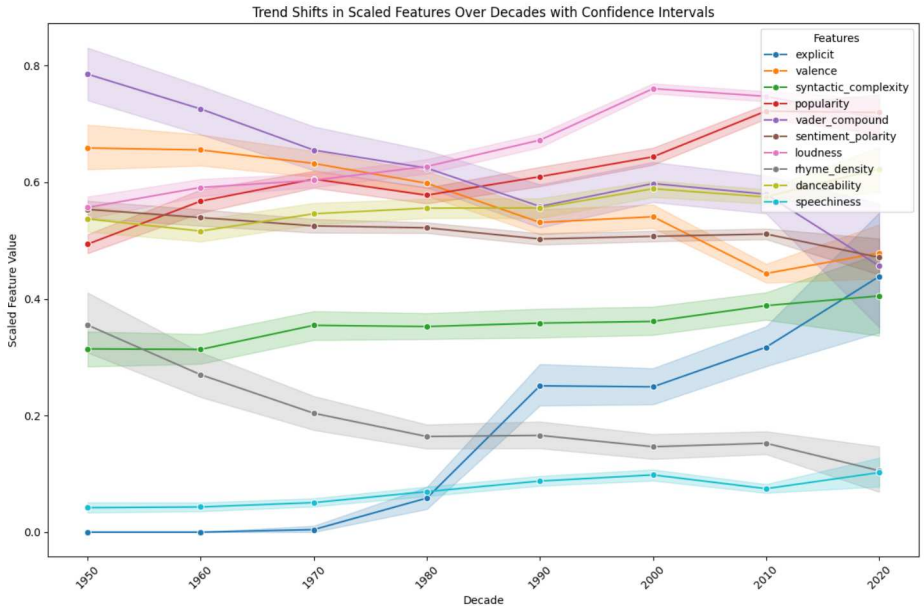

Figure 6.25: Plot showing temporal changes of features dependent on the decade in which the song was released.

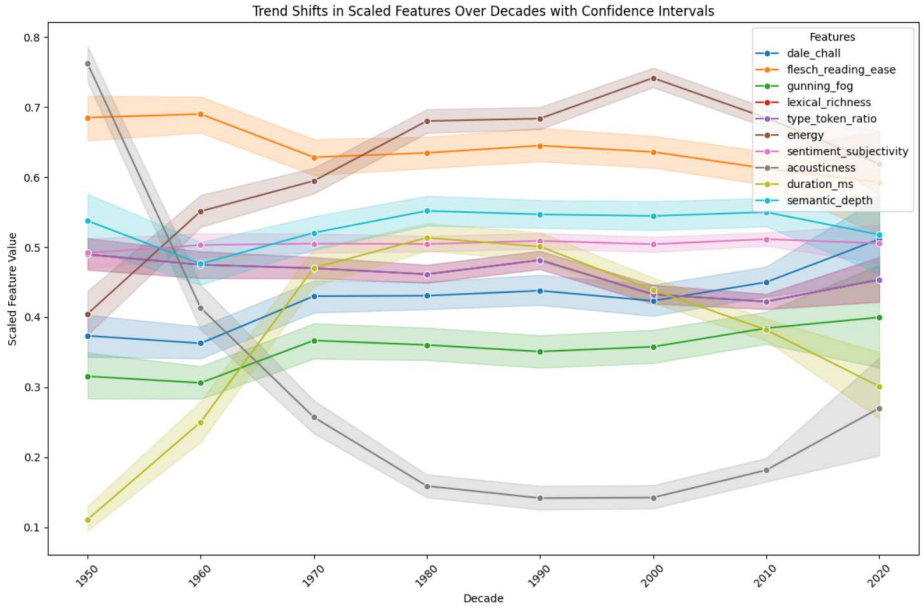

Figure 6.26: Plot showing temporal changes of features dependent on the decade in which the song was released.

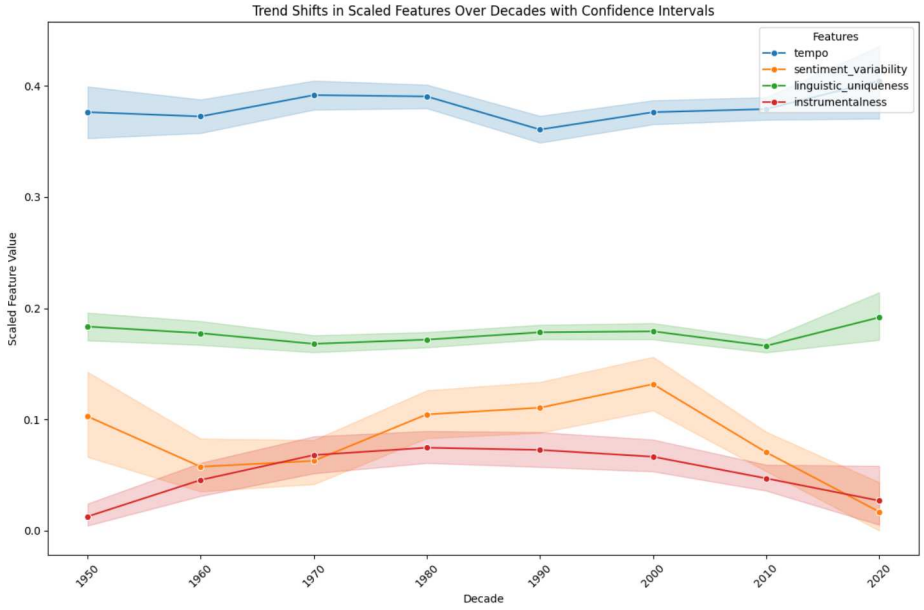

Figure 6.27: Plot showing temporal changes of features dependent on the decade in which the song was released.

Based on those plots it can be observed that:

- **There is a clear, strong upward trend in explicit content frrom the 1980s onward**. It reflects a cultural shift towards more openness in lyrical expression and increase in popularity of genres such as rap.
- **There is a steady increase in popularity over the decades. Potentially it might reflect the increased emphasis on producing music for mass appeal.**
- Speechiness peaks in the 1990s and 2000s, likely reflecting the rise of rap and spoken-word music during this period.
- The observed decrease in vader compound and valence suggests a shift towards less positive or more emotionally nuanced lyrical content over the decades.
- **There is a substantial decrease in acousticness and a corresponding increase in energy over the decades, reflecting a clear shift in musical trends. This transition highlights the move away from traditional, acoustic-oriented compositions to electronically produced music, driven by the rise of genres like EDM and modern pop. The increase in energy further underscores the growing preference for high-tempo, dynamic, and engaging tracks.**
- Loudness exhibited a consistent increase over the decades, likely also influenced by the shift toward electronic music, which often employs amplified and dynamically intense production techniques.

7. Conclusion and Future Work

7.1 Summary of Contributions

7.2 Recommendations for Future Work