Harmonic Insights: A Multifaceted Analysis of Music Data

Pranav Yeola
Goergen Institute for Data Science
University of Rochester
Rochester, USA
pyeola@ur.rochester.edu

Manasvi Patwa
Goergen Institute for Data Science
University of Rochester
Rochester, USA
mpatwa@ur.rochester.edu

***Abstract*—In this study, a large music dataset including seven genres and release dates from 1950 to 2019 is analyzed. For every song in the dataset, there are topic classifications and audio attributes. Time series analysis, sentiment analysis, correlation research, and exploratory data analysis are all included in the study. The study further explores the impact of lyrical content on classification models, including multinomial naive bayes and neural networks.**

***Index Terms*—music data, frequent itemsets mining, classification, lyrics analysis, genre analysis, sentiment analysis**

I. Introduction

Music has transformed over the years, and with digital tools, we can now dig into its details like never before. Our study dives into a diverse dataset covering seven music genres from 1950 to 2019. The data was gathered using technology like the Echo Nest® API, spotipy Python's package, and Lyrics Genius® API to get information on how songs sound and what they talk about [1].

Starting with exploratory data analysis, we look at how different genres talk about various topics. Then, we explore how different musical elements are connected and how lyrics reflect emotions. We use time series analysis to see how genres have changed over decades.Further, we also used advanced tools like multinomial naive bayes and neural networks to predict topics, genres, and even potential artists based on lyrics.

II. Related Work

A. *"A model-based approach to Spotify data analysis: a Beta GLMM"*

This study [2] explores the complex link between metrics used to measure song popularity and audio attributes taken from the Spotify database. The research utilizes a model-based methodology, more precisely a Beta GLMM (Generalized Linear Mixed Model), to decipher the complex relationships between song popularity indicators and acoustic attributes. The investigation delves into the fields of machine learning and data mining, illuminating the viability of forecasting a song's level of popularity from its audio signal. The results highlight how important it is to comprehend how a song's intrinsic qualities—as recorded in the Spotify database—affect its overall level of popularity.

B. *"Visualization of the Relationship between Metadata and Acoustic Feature Values of Song Collections"*

This study [3] focuses on visualizing the complex relationship between metadata and acoustic feature values inside song collections. The study probably looks at how metadata—like artist or genre—affects and interacts with a song's acoustic qualities. In keeping with the core of music data analysis, this visual method sheds light on the intricate interactions between descriptive data and the underlying musical elements.

C. *"Classification of Explicit Music Content Using Lyrics and Music Metadata"*

The paper [4], presents a study on the application of machine learning to classify explicit music content. The research explores the use of different data sets containing lyrics and music metadata, along with various vectorization methods and algorithms, including Support Vector Machine, Random Forest, k-Nearest Neighbor, and Multinomial Naive Bayes. The study evaluates 32 different configurations and demonstrates that the configuration with the lyric data set, TF-IDF vectorization, and Random Forest algorithm outperforms all other configurations. The paper provides insights into the use of machine learning for classifying explicit music content based on lyrics and music metadata, and it offers valuable findings on the effectiveness of different approaches and algorithms.

III. Methodology

In this research, we examine an extensive music dataset that spans seven genres (pop, rock, reggae, jazz, blues, hip hop, country, and pop) and release dates from 1950 to 2019. The dataset contains subject classifications and audio properties for every song, which were collected by integrating the Echo Nest® API with the spotipy Python audio features module by them [1]. Acousticness signifies the presence of acoustic instruments in a composition, providing a measure of its acoustic elements. Danceability, on the other hand, gauges the track's suitability for dancing by evaluating a combination of musical elements. Loudness is quantified as the average loudness across the entire track. Instrumentality reflects the degree to which a track contains fewer vocals; a higher value indicates a more instrumental composition. Valence, with high or low values, signifies the emotional tone of the track, indicating whether it conveys happiness and euphoria or sadness and anger. Lastly, Energy measures the intensity and activity of the music. Together, these audio features provide a nuanced understanding of the diverse elements that contribute to the overall musical experience.Our approach includes exploratory data analysis, in which we look into the ways in which different genres express diverse subjects, and a study into the relationships between musical features and emotions expressed in lyrics. Time series analysis is used to determine how a genre has changed over several decades. Additionally, this research explores how lyrical content affects classification models using neural networks and multinomial naive bayes classifier.

IV. Analysis

A. Exploratory Data Analysis

Our initial exploration involves an in-depth analysis of the dataset.

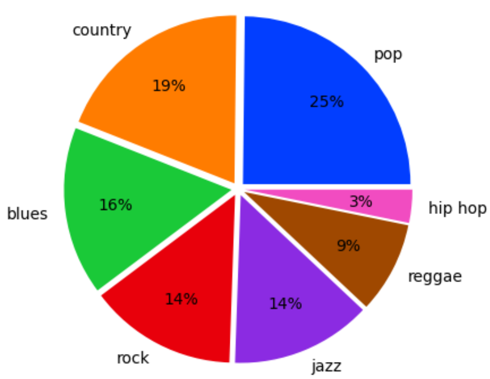


Fig. 1: Distribution of Genre By Percentage

Figure 1 illustrates the percentage distribution of music genres within the dataset, offering insights into the prevalence of each genre. Pop genre has the highest percentage, followed by country, blues, rock, jazz, reggae, and then hip hop.

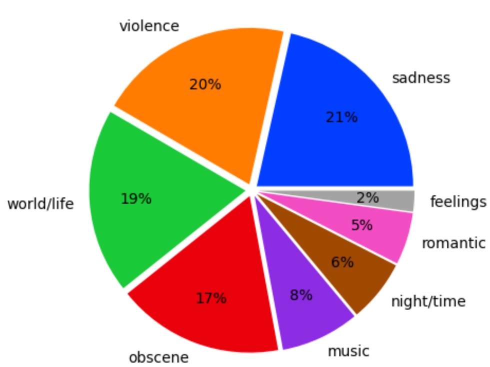
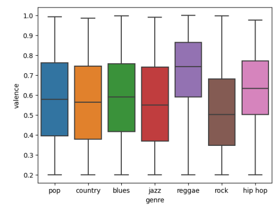


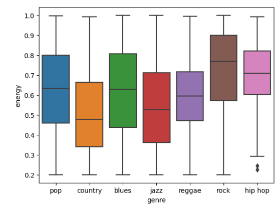
Fig. 2: Distribution of Topic By Percentage

The distribution of topics by percentage is shown in Figure 2, provides insights into the thematic content of the wide range of musical genres examined in the study. Most songs have sad theme, followed by violence, world/life and so on.

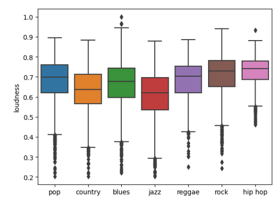
Figure 3 shows the comparison of audio features in each genre. To obtain significant values, we kept a threshold value of 0.2, since the values of these features are highly skewed towards insignificant values that are close to zero.



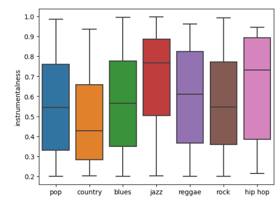
(a) Valence



(b) Energy



(c) Loudness



(d) Instrumentalness

Fig. 3: Distribution of Audio Features by Genre

Rock genre has the highest energy values, followed by hip hop genre. Pop and Blues have approximately same energy levels. Country has the least energy level. Reggae has highest valence value. Rock, Pop, Hip hop have almost same higher value of loudness. Jazz and Hip hop seem to have high instrumental composition.

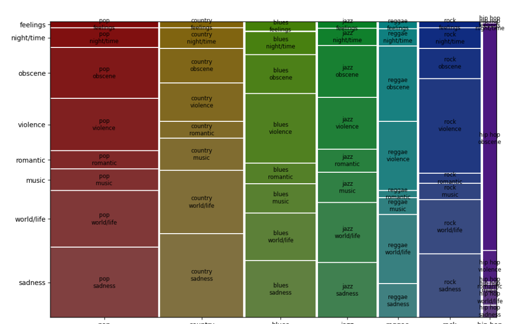


Fig. 4: Contribution of genre and topic

Figure 4 represents the portion of topics for each genre in the dataset.

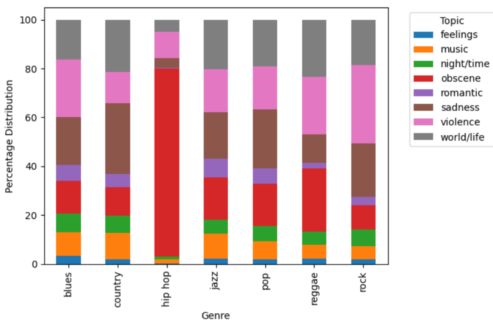


Fig. 5: Percentage Distribution of Topics in Genre

We can examine how hip hop adheres to obscene themes by looking at Figure 5. Country tends to be sadder. Rock has an element of violence. There is hardly a romantic topic in any genre.

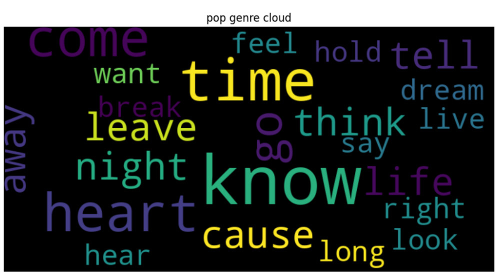


Fig. 6: Frequent Words in Pop Genre

To uncover prevalent words within each genre, we employed the FP-Growth algorithm for frequent itemset mining. This technique allowed us to identify frequently occurring words in the lyrics of songs belonging to distinct genres. This will help artists know which words are prevalent in each genre. Subsequently, we used the output from FP-Growth to construct genre-specific word clouds, visually representing the most recurrent words within each musical category. This innovative method offered a nuanced viewpoint on the unique linguistic themes that appear frequently in the lyrics across various musical genres.

B. Correlation Studies

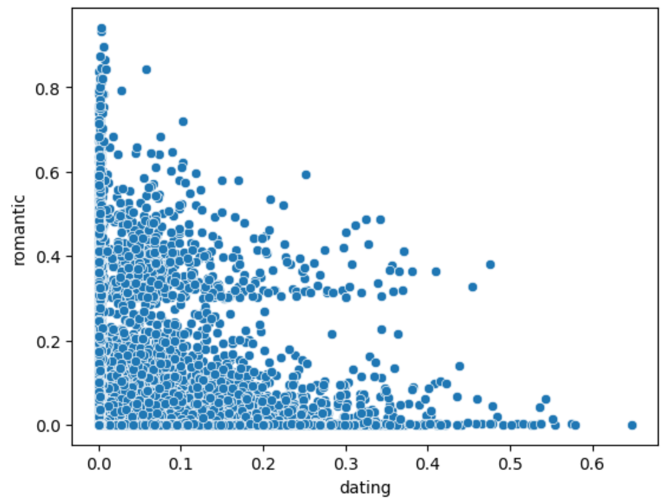


Fig. 7: Pairwise plot of Dating and Romantic

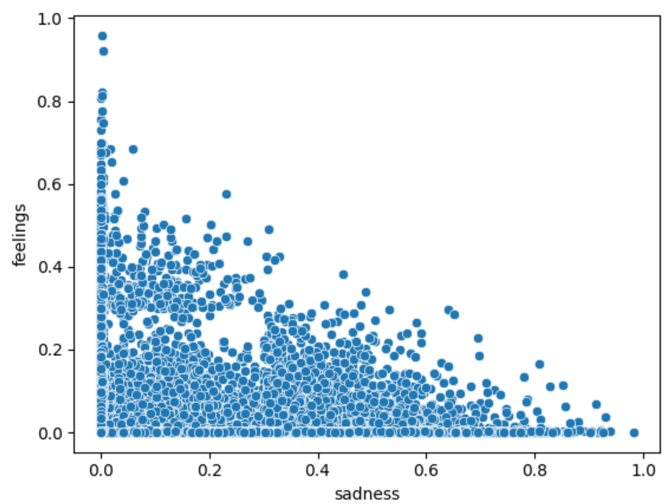


Fig. 8: Pairwise plot of Sadness and Feelings

The correlation between various values is depicted in the above image, and it is discovered that factors that would seem to be related, such as loudness, obscene, and violence, do not correlate.

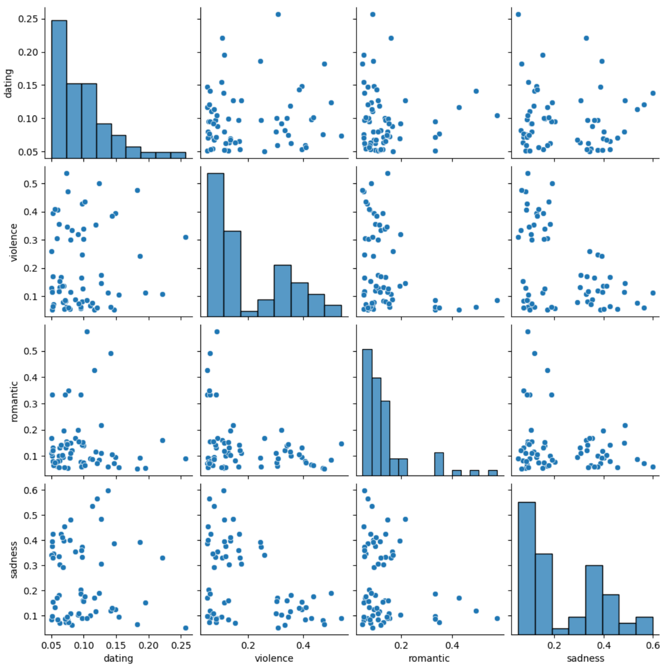


Fig. 9: Pairwise plot of Audio Features

This correlation is plotted by sampling 200 values from our dataset.

C. Sentiment Analysis

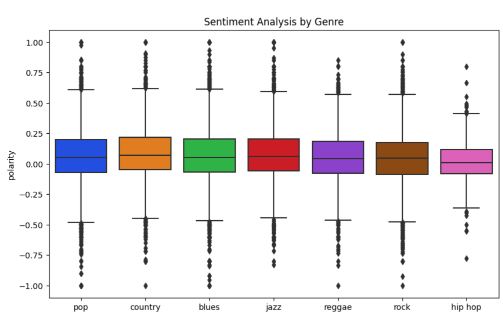


Fig. 10: Sentiment Analysis by Genre

Most of the genres have a neutral sentiment. We defined 5 sentiments, positive, mild positive, neutral, mild negative and negative. Following are their frequencies in the dataset:

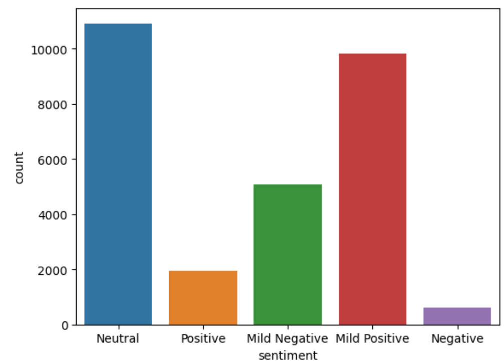


Fig. 11: Sentiments Frequency

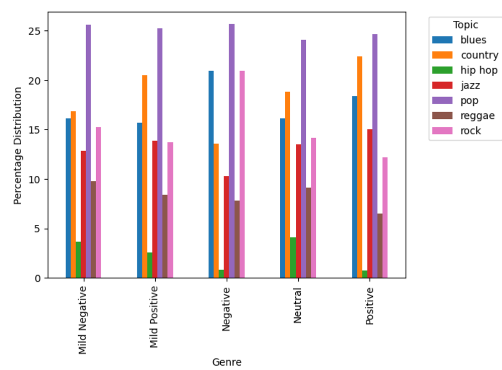
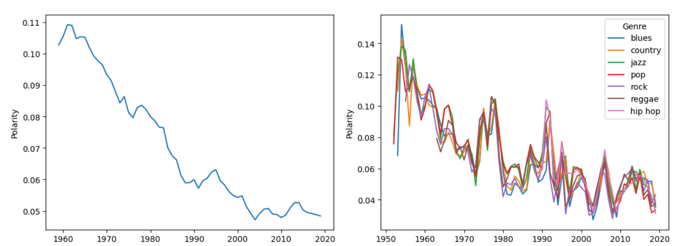


Fig. 12: Distribution of Genres across each sentiment

Compared to other genres, hip hop seems to have the least positive sentiment. Pop has the most unbiased attitude.

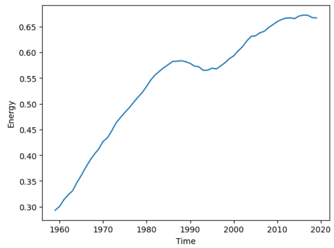
D. Time Series Analysis

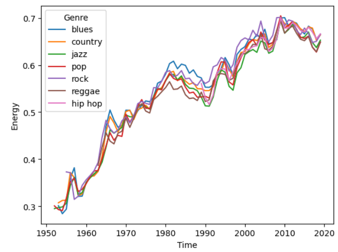


(a) Time Series of Overall Sentiments

(b) Time Series of Sentiments for each Genre

From above figure, we can see that with time the songs are getting closer to neutral sentiments.

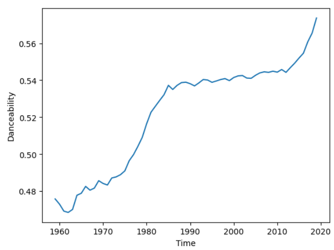


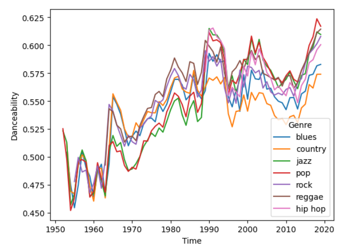


(a) Time Series of Overall Energy Levels

(b) Time Series of Energy Levels for each Genre

From above figure, we can see that with time the energy levels in the songs are getting increased.

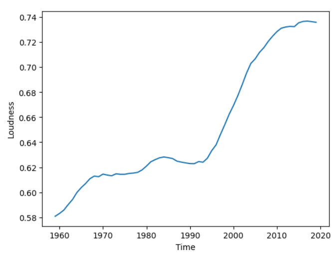


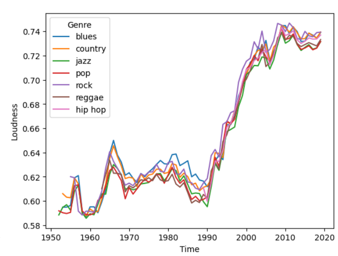


(a) Time Series of Overall Danceability

(b) Time Series of Danceability for each Genre

From above figure, we can see that with time the danceability feature in the songs is getting increased.

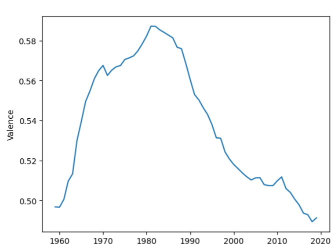


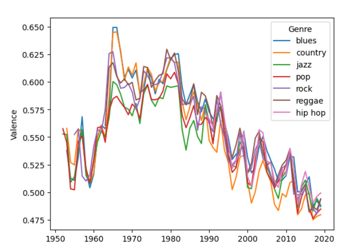


(a) Time Series of Overall Loudness

(b) Time Series of Loudness for each Genre

From above figure, we can see that with time the loudness feature in the songs is getting increased.

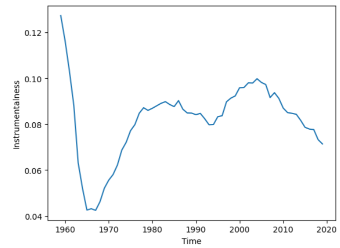


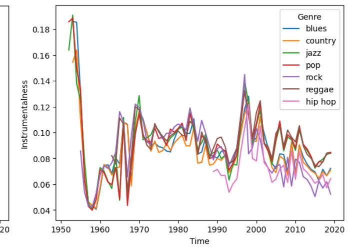


(a) Time Series of Overall Valence

(b) Time Series of Valence for each Genre

From above figure, we can see that valence peaked 1980s meaning that the songs were really happy and euphoric and then later the songs diverted towards sentiments of saddness and anger.





(a) Time Series of Overall Instrumentalness

(b) Time Series of Instrumentalness for each Genre

From above figure, we can see that instrumentalness dropped in late 1950s meaning that the songs didn't have much instrumental composition then but later that increased.

E. *Model Performance Evaluation*

TABLE I: Comparison of Model Performance

Model	**Train Accuracy**	**Test Accuracy**
k-NN	59.25%	29.66%
Logistic Regression	37.53%	36.83%
Gaussian Naive Bayes	33%	32%
SVM	41%	39%
Random Forest Classifier	48%	40%
Neural Network	78.20%	33%

The table presents a comparison of various machine learning models based on their training and testing accuracies. The k-NN model exhibits a relatively high training accuracy of 59.25%, but a lower testing accuracy of 29.66%, indicating potential overfitting. Logistic Regression shows moderate performance with 37.53% training accuracy and slightly better testing accuracy at 36.83%. Gaussian Naive Bayes and SVM demonstrate comparable training and testing accuracies at around 33% and 32%, and 41% and 39%, respectively. Random Forest Classifier exhibits a training accuracy of 48% and a testing accuracy of 40%. Notably, the Neural Network stands out with a high training accuracy of 78.20%, yet its testing accuracy is relatively lower at 33%, suggesting a need for further evaluation and potential fine-tuning to enhance generalization.

F. *Multinomial Naive Bayes Classification Models*

1) Topic Classification from Lyrics: The Naive Bayes model for topic classification from lyrics demonstrates strong overall performance, achieving an accuracy of 0.6738. The precision score of 0.6286 suggests a reliable ability to correctly identify relevant topics, while the recall score of 0.6738 indicates a high rate of correctly capturing all actual relevant topics. The F1 score, considering both precision and recall, stands at 0.6036, showcasing a balanced trade-off between precision and recall in the model.

Metric	Value
Accuracy	0.6738
Precision	0.6286
Recall	0.6738
F1 Score	0.6036

TABLE II: Performance Metrics for Topic Classification from Lyrics using Naive Bayes

2) Genre Classification from Lyrics: In the case of genre classification from lyrics, the Naive Bayes model exhibits limited success, with an accuracy of 0.3003. The precision score of 0.4553 indicates a moderate ability to correctly identify specific genres. However, the recall score of 0.3003 suggests challenges in capturing all instances of the actual genres present. The F1 score, harmonizing precision and recall, is relatively low at 0.1988, implying the need for more data.

Metric	Value
Accuracy	0.3003
Precision	0.4553
Recall	0.3003
F1 Score	0.1988

TABLE III: Performance Metrics for Genre Classification from Lyrics using Naive Bayes

V. Conclusion

In conclusion, our analysis of a diverse music dataset spanning seven genres offered valuable insights into the intricate relationship between lyrical content and audio features. Through sentiment analysis, time series trends, and classification models, we identified distinctive patterns within genres and observed their evolving emotional tones over decades. This project contributes to a deeper understanding of music data analysis, suggesting potential future avenues such as the development of recommendation models and further refinement of classification models.

References

[1] Moura, Luan; Fontelles, Emanuel; Sampaio, Vinicius; França, Mardônio (2020), "Music Dataset: Lyrics and Metadata from 1950 to 2019", Mendeley Data, V2, doi: 10.17632/3t9vbwxgr5.2

[2] Sciandra M, Spera IC. A model-based approach to Spotify data analysis: a Beta GLMM. J Appl Stat. 2020 Aug 10;49(1):214-229. doi: 10.1080/02664763.2020.1803810. PMID: 35707796; PMCID: PMC9042099.

[3] M. Watanabe, N. Kuroko, H. Ohya and T. Itoh, "Visualization of the Relationship between Metadata and Acoustic Feature Values of Song Collections," 2022 Nicograph International (NicoInt), Tokyo, Japan, 2022, pp. 79-84, doi: 10.1109/NicoInt55861.2022.00023.

[4] Egivenia Egivenia, Gabriella Ryanie Setiawan, Stella Shania Mintara, and Derwin Suhartono. 2021. Classification of Explicit Music Content Based on Lyrics, Music Metadata, and User Annotation. In Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology (SIET '21). Association for Computing Machinery, New York, NY, USA, 265–270. https://doi.org/10.1145/3479645.3479695