**Exploring Music Through Textual Insights**

**ABSTRACT**

The entertainment industry is rapidly growing, evidenced by the emergence of new musical genres and the continuous influx of artists. Daily, vast quantities of music are produced and released. For music streaming platforms, categorizing this music by genre, and recommending tracks to users are essential tasks. To address these challenges, various artificial intelligence techniques have been developed. However, one significant hurdle in developing effective machine learning algorithms is the scarcity of adequate training data. Currently, an immense volume of musical knowledge is documented in written form, with records dating back several centuries. In this study, we explore various Natural Language Processing (NLP) methods to leverage these extensive text collections for automatic music knowledge discovery. Our approach encompasses several stages of a typical NLP pipeline, including corpus compilation, text mining, information extraction, named entities, sentiment analysis and Grid Search to optimize the model using Logistic Regression. We present these methods within different musical/genre contexts such as Hip pop, Rock, Pop… etc where large document collections are analyzed. The findings from these data-driven analyses are then discussed and elucidated.

On the data of 79 Musical Genre Classification dataset, this project constructed few ML models and used a data filtering strategy. This study does a comparative analysis and discusses the results. The models developed and evaluated are Support Vector Machine, Cross validation Grid search using Logistic Regression. Best Score gained by Logistic Regression is 48%, and by Linear SVC is 74%. This project’s research outcomes will help us to analyze music in different areas.

***Keywords***: Machine Learning, Classification, Natural Language Processing, Pre-processing, Text mining, Entity Linking (named entities), Information Extraction, Sentiment Analysis, Music, Grid Search.etc

1. **Introduction**

The genre is arguably the most critical and recognizable aspect of any song. Many music enthusiasts curate playlists based on genres, and various music streaming platforms like Apple Music, Spotify, and Wynk tailor song recommendations based on users' preferred genres. As the music industry continues to evolve rapidly, establishing new song categorizations can be challenging, but technology offers solutions to these challenges. In today's digital era, music transcends mere melodies and rhythms; it encapsulates emotions, stories, and critiques, playing a significant role in shaping our feelings, influencing us, and bridging cultural divides.

In this project, we aim to delve into music through the prism of textual data. Numerous initiatives have predicted song genres using textual descriptions, artist biographies, or song reviews. Music acts as a powerful medium of expression, resonating deeply with audiences and transcending cultural barriers. Our exploration focuses on analyzing music through textual data analysis.

Instead of using deep learning techniques such as CNNs and RNNs, we have implemented SVM and optimized models using grid search and logistic regression. By leveraging advanced Natural Language Processing (NLP) techniques, we aim to dissect the intricate relationships between music genres, artists, and user sentiments. Through this endeavor, we aspire to unearth hidden patterns, sentiments, and connections within the expansive musical landscape, providing insights that resonate with music lovers and industry experts alike.

**Our Project is divided into two major parts.**

1. **Preprocessing and identifying the popular music genre by KNN (***so that we can use this to extract important entities such as lyric titles/ artist names and song names***).**
2. **Named Entity Recognition and Sentiment Analysis.**

The **problem being addressed** is the need to analyze and understand the underlying connections between music genres, artists, and user sentiments.

Employing Natural Language Processing (NLP) techniques, we successfully identified prevalent themes and sentiments associated with specific songs and genres. This analysis enabled us to determine which types of artists are most suitable for musical styles. The findings of this project facilitate a deeper understanding of the interplay between musical content and audience perception, providing valuable insights for tailoring artists and genre recommendations.

* 1. **Data Collection, Preprocessing and Feature Extraction, Genre Prediction**:
* This Project Begins with the compilation and preprocessing of diverse dataset comprising Song metadata, artist information, and user ratings. After cleaning and standardizing textual data, performing preprocessing operations like tokenization, stemming and Lemmatization, we observed that musical genres are usually distinguished from each other in terms of their musical content.
* However, we are aware about the lyrical content that is often associated with specific genre.
* Therefore, we applied some magic here to extract genres list, by focusing on the top 15 most frequent themes. Below is the outcome corresponding to this stage.

A screenshot of a white grid

Description automatically generated

A screenshot of a computer

Description automatically generated

Secondly, applying the model KNN to get the top 10 nearest songs.

**Steps how this algorithm works:**

* Randomly selects a song from the cleaned dataset and displays the details about it including the song name, artist Id and lyrics. So then it prints out which song (by which artist) the results will be used for providing the context for the recommendations.
* In here we created a sparse matrix of the song features using TF-IDF and chosen a song index and model\_knn(the trained KNN model).
* It returns a DataFrame containing recommended songs along with their names, artist IDs, lyrics, and a 'distances' column that likely indicates how similar each recommended song is to the selected song based on the model's distance metric.

A screenshot of a computer

Description automatically generated

* Next comes the outputs of the recommendations, in here the detailed information about the top recommendation, including the song name , artist and overview of the lyrics. Like what is the most similar song according to the model.
* The Python code snippet you've provided outlines a content-based recommendation system that employs the k-nearest neighbors (KNN) algorithm to recommend songs based on similarities in textual features extracted from song lyrics. Let's break down the major components and their functionality:
* Next, where 50 random songs are selected, and for each, recommendations are generated. It accumulates themes from the recommendations to find the most frequent themes. It also counts how often each song is recommended across these trials and identifies the top 5 most recommended songs, summarizing the results to understand the model's tendency in recommendations.
* Lastly, after running recommendations multiple times, **it analyzes the most common themes across all recommended songs**, providing insights into what genres or topics are most prevalent in songs recommended by this system.

**Output of the first Part:**

A screenshot of a computer

Description automatically generated

* 1. **Named Entity Recognition and Sentiment Analysis:**
* The collected data above is used here, which includes lyrics and artist information with their popularity.
* We have extracted and filtered entities from the lyrics. It filters these entities to include only those classifies PERSON, ORG and PRODUCT, which could represent artist names, organizations, or album/song titles.
* The named entities are extracted from the first 200 songs lyrics, then it compiles a list of all entity types and texts across all songs counting their occurrences to analyze the most common entities mentioned in the lyrics.

A close up of words

Description automatically generatedA graph with a bar and a number of numbers

Description automatically generated with medium confidence

* For **Sentiment Analysis**, we have grouped artists with their ratings corresponding to their popularity scores, also we observed how many songs each artist has in each sentiment category.
* This grouped data is then visualized showing the distribution of sentiments for the top 10 artists.
* The chart indicates which artists are associated with the most “liked” songs and contrasts this with those that are “least liked” or considered “not bad”.

A graph of a rating

Description automatically generated with medium confidenceA graph showing different colored squares

Description automatically generated

**Model Optimizations and Evaluations:**

* We created a pipeline and converted the raw lyrics into TF-IDF features, which are then fed into the logistic regression model.
* So, the **Formula of TF-IDF** is as follows:
* The number of parameters combination in the parameter\_grid and
* Number of folds used in cross-validation(cv)
* Number of Combinations of hyperparameters:

(options for tfidf\_ngram\_range) \*(options for tfidf\_max\_features)\*(options for clf\_c) \* (options for clf\_penalty)

2\*2\*3\*2 = 24 ocmbinations.

* Cross validation folds:

We gave Cv=5 then each combinationis tested 5 times (that is each fold involves training on 4/5ths of the data and validating on 1/5th)

The total number of fits is now 120 fits.

* That is the model will initiate 120 individual training and evaluation cycles.
* Now after above hyperparameter tuning the best score we got was 48%.
* Then we split the data into training and test sets an then TF-IDF vectorization is applied
* A Linear SVC is trained on TF-IDF vectors of the training set and evaluated on test set
* We got approximately 74.8% of accuracy, which is the proportion of ht ecorrest predictions out of all predictions made by the model.

**Conclusion:**

This Project demonstrates the efficacy of natural language processing (NLP) in discovering patterns in text data by analyzing lyrical content and matching comparable songs using a KNN-based recommendation system. Through the extraction of named entities from lyrics, it provides a visual representation of these relationships and emotional responses, providing insights into recurring themes and their relationship to artist popularity. By emphasizing the complex relationship between an artist's work and audience reaction, this approach paves the way for more sophisticated applications like sophisticated recommendation engines and prediction models for artist popularity. NLP analysis demystifies feelings connected to various musicians by illuminating the emotional terrain surrounding musical compositions.

Although the results are only moderately accurate, they have important ramifications for music recommendation engines and offer a better comprehension of the emotional context of music. Subsequent improvements can include adding complex algorithms, improving feature extraction, and adjusting hyperparameters. Furthermore, a more thorough evaluation of the model might be provided by using a wider range of performance indicators, especially when assessing datasets with inconsistent sentiment representation. To sum up, this initiative is a step towards developing a more personalized and emotionally aware method of music recommendation, which will improve user experiences by matching recommendations to the complex preferences and feelings of its user base.

**References:**

[1] Tham, I. (2021, August 25). Generating Music using Deep Learning Techniques. Towards Data Science. Retrieved from <https://towardsdatascience.com/generating-music-using-deep-learning-cb5843a9d55e>

[2] Islam, M. S., Hasan, M., Rahim, M. A., & Hasan, A. M. (2021, December). Machine learning based Music Genre Classification with Pre-Processed Feature Analysis. ResearchGate. Retrieved from <https://www.researchgate.net/publication/356585248_Machine_Learning-Based_Music_Genre_Classification_with_Pre-Processed_Feature_Analysis>

[3] Mhatre, K., & Almeida, V. (2020). Music Genre Classification using Neural Networks. Retrieved from <https://www.semanticscholar.org/paper/Music-Genre-Classification-using-Neural-Networks-Mhatre-Almeida/3ec6a47f9af404ce34686364e8da6f11e8563d4c>

[4] Kovachev, D. (2019, April 1). How we used NLTK and NLP to predict a Song’s Genre From its Lyrics. Towards Data Science. Retrieved from <https://towardsdatascience.com/how-we-used-nltk-and-nlp-to-predict-a-songs-genre-from-its-lyrics-54e338ded537>

[5] Oramas, S., Espinosa-Anke, L., Gomez, F., & Serra, X. (2018, July). Natural Language Processing for Music Knowledge Discovery. Retrieved from <https://www.researchgate.net/publication/326217792_Natural_Language_Processing_for_Music_Knowledge_Discovery>

[6] Costs, Y. M. G., Oliveria, L. S., Koerich, A. L., Gouyon, F., & Martins, J. G. (2012, May 17). Music Genre Classification using LBP Textural Features. Retrieved from <https://repositorio.inesctec.pt/server/api/core/bitstreams/9757cd52-b6ec-477e-a994-add6756847f9/content>

[7] Neisse, O. (n.d.). Scrapped Lyrics from 6 Genres [Data set]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres>