A Project in CS584 - Natural Language Processing

Application of NLP Techniques on Song Lyric Data

Group Members:

Atharv Subhekar, Nidhi Sura, Nikhil Shah

I. Introduction

The exploration of music through the lens of its lyrics offers a rich avenue to delve into its cultural, emotional, and linguistic facets. A project that seeks to classify and analyze music lyrics genre-wise stands at the crossroads of computational linguistics, musicology, and cultural studies. By dissecting lyrics into various genres and running a gamut of analytical metrics, we aim to unearth the distinctive characteristics and the inherent essence carried by different musical genres.

The findings from this project have a broad array of applications. They can provide music creators with insights into the hallmarks of different genres, aiding in more nuanced creation of music. Additionally, the music industry and digital streaming platforms can leverage the analysis to enhance music recommendation systems, ensuring a more tailored user experience. Moreover, from an academic standpoint, the project can contribute to the interdisciplinary study of music, linguistics, and culture, potentially leading to new theoretical frameworks or pedagogical approaches.

The challenges entailed in this project are multifold. Accurate genre classification is the first major hurdle, given the subjective nature and the frequent cross-over of musical genres. Further, oftentimes musical features are required to determine the exact genre. However, our project aims to establish a relationship between lyrics and the genre of a song. Extracting meaningful insights from the lyrical content requires sophisticated NLP and ML algorithms capable of handling the nuances and ambiguities inherent in human language. Additionally, the sheer volume of music and the diversity of languages and cultural references in lyrics pose significant challenges in data processing and analysis.

Some relevant approaches to our problem statement when conducting a literature review were:

- A paper implementing features and techniques for the classification of lyrics into music genres – such as Malmi Rhyme Factor, Flesch reading ease and principal components analysis.
- A paper outlining an approach that uses variations of LSTM implementations, including a bi-directional LSTM.
- A medium.com article that also covers a similar bidirectional LSTM approach, but with some more fundamental analysis using n-grams.

Using the aforementioned papers as a guide we have made the following progress with the project.

i. Extensive Literature Survey: A thorough read and review of each paper was performed by us in order to gain insights on the topic and the prior work put into the study of music lyrics.
ii. Feasibility Study: We gauged the extent to which prior works had been able to generalize and study music lyrics. We also looked into emerging genres to make sure our research will be effective in the near future of music.
iii. Cemented Approach: We experimented with a few different data processing techniques and NLP models to determine what will work best for our application. We finalized the approach and implemented an intermediate version of the code.
iv. Data Preprocessing Module: We performed an extraction of all the data from the csv file. Then, we executed the pre-processing necessary for the data. This included converting to lowercase, tokenization and the removal of stop words.
v. Basic Analysis and Word Cloud Analysis: We summarized the data by genre in the form of a word cloud.
vi. Vectorization Module: We vectorized the lyrics using a stacked LSTM model that found correlations between the words.

II. Problem Formulation

The core task at hand is to collect, classify, and analyze a substantial corpus of music lyrics across a spectrum of genres. Utilizing natural language processing (NLP) and machine learning (ML) techniques, the project aims to dissect lyrics into genres such as rock, pop, hip-hop, country, and jazz among others, and subject them to a meticulous analysis based on a variety of proposed metrics. The output of this analysis would provide a quantitative and qualitative understanding of the narrative, thematic, and linguistic variances across different genres.

The project is broken down into two parts consisting of three subparts each:

1. Lyric-Based Song Classification:
    a. Input: The input to this part is dataset consisting of the song name, musical artist, release date and lyrics of a large volume of songs.
    b. Algorithm: The two major sub-steps can be defined as Genre Classification and Lyrics Analysis.
    c. Output: The output shall be one or more classified genres for each song (such as rock, pop, country, jazz, etc.). We shall also be analyzing and plotting a variety of features and relationships to establish a relationship between the lyrical content of a song and the style of an artist, the typical characteristics of a genre, and the trends in a specific time period.
2. Genre-Based Lyric Generation:
    a. Input: The input to this part of the project is only the songs from the dataset that are labelled with the genre 'Country'. From these songs, 1000 random samples are selected to train the model.
    b. Algorithm: The algorithm used is an LSTM-based RNN network. The model should be trained with enough epochs for it to understand the nuances of the genre and provide a non-repetitive, somewhat coherent song.

c. Output: The output shall be the abovementioned song lyrics of the genre 'Country'. An output seed shall be used to determine a starting point for the song to generate. The output shall be generated on every epoch until a satisfactory result is obtained.

Genre Classification:

Task Type: Multi-Class Classification

Description: In this step, each song is categorized into one of the predefined genres. Multi-class classification is suitable here as there are multiple genres (e.g., rock, pop, hip-hop, country, etc.) and each song belongs to one of these genres. This is a classic case of multi-class classification where the aim is to assign a single label (genre) from multiple possible labels to each item (song).

Lyrics Analysis:

Task Type: Descriptive Analysis

Description: Once the songs are categorized, the next step is to analyze the lyrics within each genre based on various metrics like thematic density, lexical richness, sentiment analysis, etc. This step does not fall into the conventional machine learning task types like classification or regression. Instead, it's a form of descriptive analysis where various statistical and natural language processing (NLP) techniques are employed to extract insights and understand patterns within the lyrics of each genre.

Lyric Generation:

Task Type: Generative AI

Description: The last part of the project focuses on using the knowledge from the dataset and our lyrics analysis to generate song lyrics for a given genre. Due to the model training time compounding for every genre we added to the mix, we decided to pick one genre with distinct vocabulary and linguistic style - country music. Here, the song will be sequentially generated from a seed word or phrase provided by the user.

III. Methods

1. Data Preprocessing:
   i. Data Collection: We aimed to collect a substantial dataset of music lyrics along with their respective genres. We decided to choose the Song Lyrics Dataset on Kaggle which is updated monthly with new data. One of the main reasons we picked this specific dataset is the fact that it covered a diverse range of genres to ensure a comprehensive analysis. The size of the data was also significantly large, increasing the likelihood of our analysis applying to a large majority of songs.
   ii. Text Normalization: We normalized the data by converting it to lowercase and removing irrelevant characters. These irrelevant characters included special characters and punctuation.

iii. Text Processing: We removed most stop words, while retaining some others that are required to provide a contextual understanding of the lyrics. We also converted numbers to digits and removed accents from characters.
iv. Tokenization: We tokenized the string of lyrics to a list of words separated on the space between them. These tokens served as an input to the neural network model.

2. Model Design:
   i. Genre Classification Model: We employed a multi-class classification model using an LSTM architecture to categorize the songs into respective genres based on their lyrics.
   ii. Lyric Generation Model: We designed a model to generate lyrics for a given genre and seed using an LSTM-based RNN model.
   iii. Loss Function Design: For the genre classification task, a categorical cross-entropy loss function is suitable as it is designed for multi-class classification problems.
3. Training:
   i. Training Data Split: Split the collected data into training, validation, and test sets to ensure a robust evaluation of the model's performance.
   ii. Model Training: Train the genre classification model using the training set, while monitoring the performance on the validation set to prevent overfitting. We used 5 epochs to train, trying to avoid overfitting
4. Classification Model:
   i. Embedding Layer: The model starts with an embedding layer. If a pre-trained weight matrix (such as GloVe embeddings) is provided, it uses that for embedding; otherwise, it initializes a new embedding layer with a specified vocabulary size and embedding dimension.
   ii. LSTM Layer: Following the embedding layer, the model employs an LSTM (Long Short-Term Memory) layer. LSTM is useful for processing sequences (like text) as it can capture long-range dependencies and contextual information. The LSTM layer has configurable parameters like hidden dimensions, number of layers, and dropout probability.
   iii. Attention Mechanism: The model includes a custom attention mechanism (Attention class) which computes attention scores based on the LSTM's outputs. This mechanism allows the model to focus on specific parts of the input sequence when making predictions, which is particularly useful for understanding the context and nuances in text data.
   iv. Dropout and Fully Connected Layer: After the attention mechanism, the model applies dropout for regularization, followed by a fully connected linear layer to map the LSTM outputs to the desired output size.
   v. Output: The final output is designed to represent the sentiment of the input text. The model can be used for binary classification (positive/negative sentiment) or multi-class sentiment classification, depending on the specified output size.
5. Lyric Generation Model:
   i. Architecture: LSTM-based RNN model (Sequential-LSTM-Dense)
6. Inference:
   i. Genre Classification: Once trained, use the classification model to categorize the lyrics into genres.
   ii. Lyric Generation: After generating song lyrics, use the training and validation accuracy to evaluate whether the model has fitted well on the data. Compare and

contrast the output across epochs and use metrics such as ROUGE to determine how good the text prediction is.

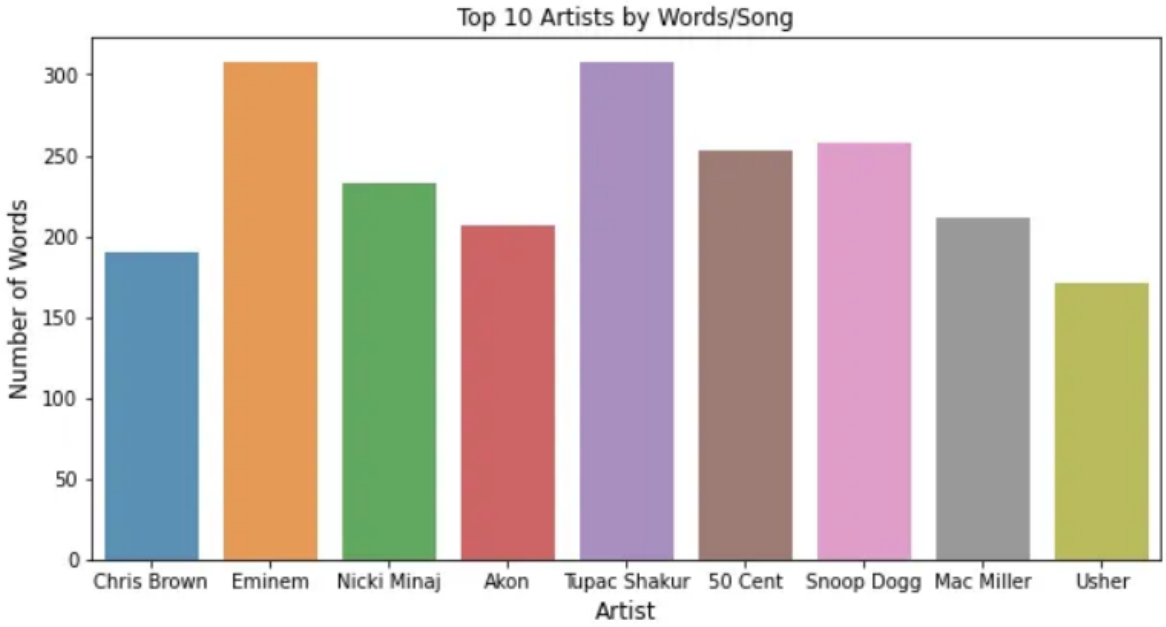
IV. Dataset and Experiments

The dataset we are using for the project is called Song Lyrics. It is hosted on Kaggle and is composed of two datasets artists-data.csv and lyrics-data.csv.

Here is a sample of a few rows of the artists-data.csv file.

Artist	Genres	Songs	Popularity	Link
Will Smith	Black Music; Hip Hop; Rap	130	0	/will-smith/
Pregador Luo	Gospel/Religioso; Hip Hop; Black Music	147	1	/pregador-luo/
Jennifer Hudson	R&B; Black Music; Pop	89	0	/jennifer-hudson/
T.I.	Hip Hop; Rap; Black Music	243	1	/t-i/
Raiz Coral	Gospel/Religioso; Black Music; Soul Music	57	0	/raiz-coral/

Data Visualization:

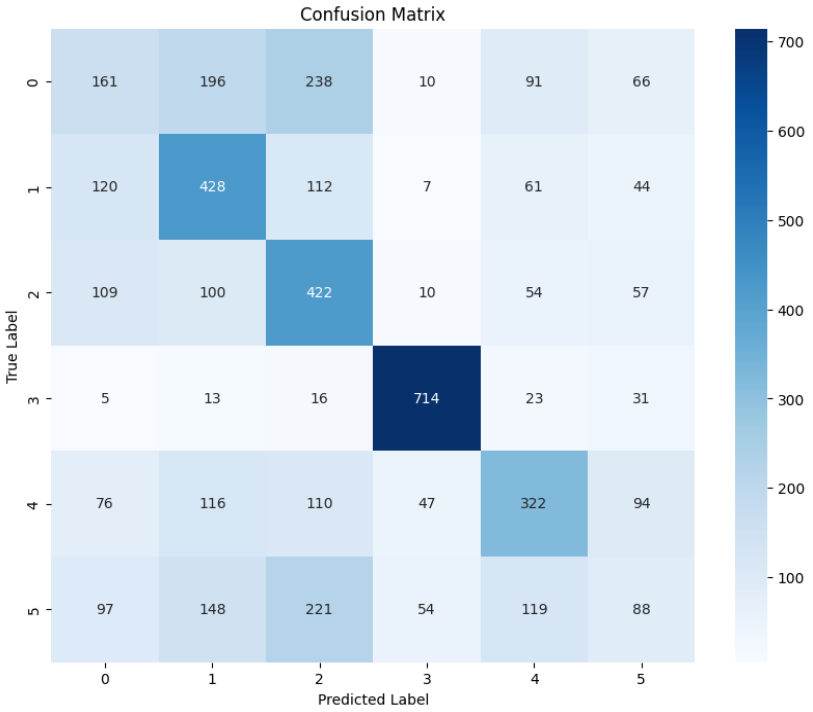


A count of all the songs in the dataset arranged by class (genre)

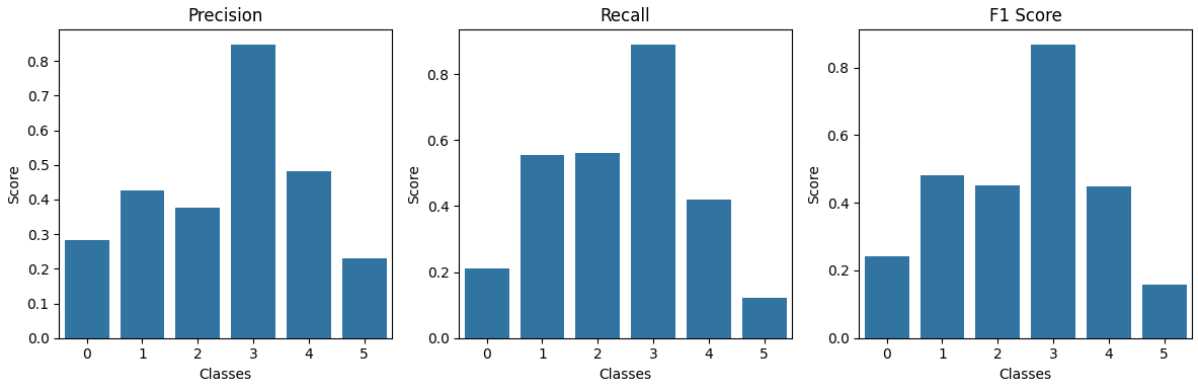
Word clouds for genres (i) Pop, and (ii) Rock

V. Results:

Genre Classification Performance:



Confusion matrix



Precision, Recall & F1 Score for predicted classes

Classes:

Rock: 0, Country: 1, Indie: 2, Rap; Hip Hop:3, Pop:4, Other: 5

Accuracy: 0.4661572052401747

Lyric Generation Performance:

The rouge scores were obtained by comparing a five-word long seed stemming from the a lyric in the dataset. The lyric was treated as reference text, and the ROUGE scores were obtained as follows:

ROUGE-1 Score: 0.08450704047609604,

ROUGE-2 Score: 0.020408161432736528,

ROUGE-L Score: 0.08450704047609604

The output obtained after 20 training epochs is given below:

Small town boy	Seed
man im gon na see nobody	Verse begins
thats gon na break rusty cage	
run run back back back	
tennessee homesick	
midnight homesick blues	
mama aint got ta go back	Verse ends
way back tennessee	Prechorus begins
i´m goin back	Prechorus Ends
tennessee homesick blues	Chorus begins
got ta kick back relax	
tennessee homesick blues	
got ta kick back relax	Chorus ends

VI. Conclusions

Genre Classification:

The project aimed to develop a machine-learning model capable of classifying song lyrics into different musical genres. Utilizing a deep learning approach, specifically a Long Short-Term

Memory (LSTM) network with an attention mechanism, the model was trained on a dataset where lyrics were categorized into six genres: Rock, Country, Indie, Rap/Hip Hop, Pop, and Other.

Training Results:

The training process showed a gradual decrease in loss over five epochs, indicating that the model was learning and improving its ability to classify genres from the lyrics. The loss reduced from 1.7630 in the initial steps of the first epoch to 1.2383 towards the end of the fifth epoch. This consistent decrease is a positive sign, demonstrating the model's ability to learn from the training data.

Evaluation Metrics:

Upon evaluation, the model achieved a test accuracy of 46.62%, with a test loss of 1.3218. While the accuracy suggests a moderate level of prediction capability, it also highlights room for improvement. The precision, recall, and F1 scores for each class showed varied performance across different genres:

The model performed exceptionally well in the 'Rap/Hip Hop' genre, as evidenced by high precision, recall, and F1 scores.

Genres such as 'Rock', 'Other', and 'Indie' saw lower scores, indicating challenges in accurately classifying these genres.

The confusion matrix further elucidated these disparities, showing a significant number of misclassifications among certain genres.

Interpretation and Future Work:

The varied performance across genres suggests that the model, while effective in certain cases, struggles with the nuanced differences in lyrical content among some genres. This could be attributed to factors like the inherent similarity in lyrics across these genres, limited representation of certain genres in the training data, or the complexity of capturing contextual meanings in lyrics.

For future improvements, several approaches can be considered:

- Enhanced Data Preprocessing: Further cleaning and preprocessing of the data might yield better representations of each genre, aiding in more accurate classification.
- Data Augmentation: Including more diverse data for underperforming genres could help improve the model's learning in those areas.
- Model Tuning: Adjusting the model architecture, experimenting with different hyperparameters, or trying other models like Convolutional Neural Networks (CNNs) could enhance performance.

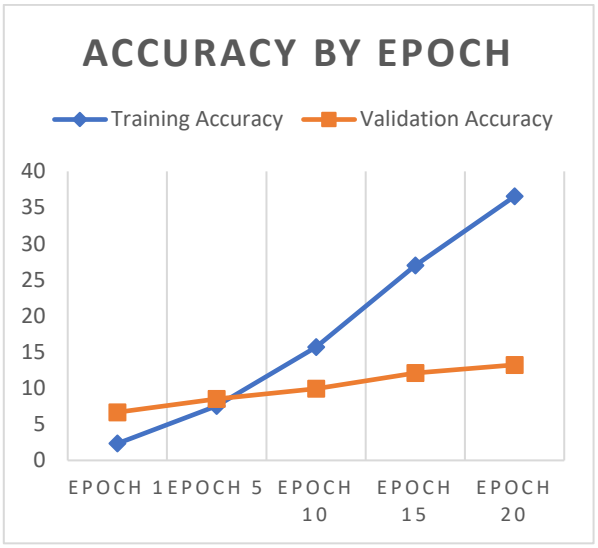
- Advanced NLP Techniques: Incorporating techniques like word embeddings fine-tuned on musical content or exploring transformer-based models could capture deeper linguistic and contextual nuances in lyrics.
- Balanced Dataset: Ensuring an equal representation of each genre in the training process could help improve the model's ability to learn from less represented genres.

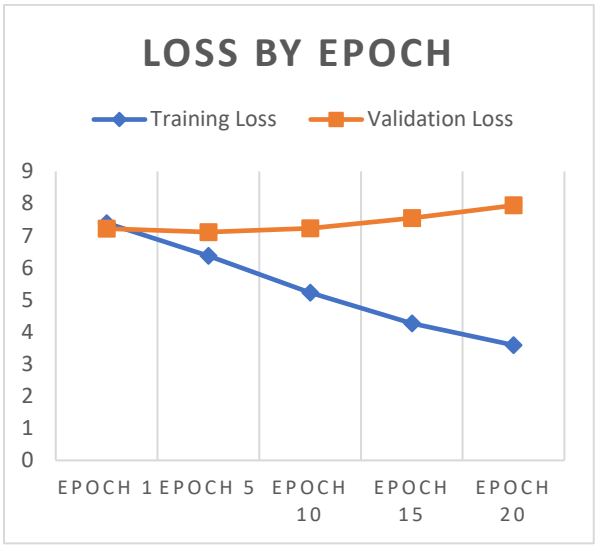
In conclusion, while the model shows promising results in certain genres, the overall performance indicates a need for further optimization and exploration of advanced techniques to improve its generalization capabilities across a broader range of musical genres.

Lyric Generation:

The ROUGE scores mentioned in the results indicate that 8.45% of the unigrams in the generated text were also present in the reference text and 2.04% of the bigrams in the generated text were also present in the reference text. There were no matching 3-grams or 4-grams present in the sentences. The Rouge-L score indicated that the longest common sequence overlap in the two texts was 8.45%.

However, it is important to note that since there is no clear reference for generated text in lyric generation, ROUGE score may not be the best metric. Instead, we can look at the validation and training accuracies and losses for each epoch in the training process.





The accuracy chart suggests that the model is fitting well on the training data, however, the validation accuracy is increasing at a slower rate which is common for ML models. The continuously increasing rising accuracy suggests that our model will continue to improve over epochs. The loss graph suggests that the training loss gradually decreases while the validation loss remains the same. This could be for a number of reasons:

- Our training and validation datasets are too small, with the training set consisting of 800 examples and the validation set consisting of 200 examples. A larger set of data should be able to fix this issue.
- The model is too simple. A 3-layer model may not be enough to train the model on fewer epochs to better fit the data. A more complex model can fix this issue but will take much more time to train.

One of the best methods to evaluate generated text is to manually understand and interpret it. On manually observing the generated text in the results section, we see that the lyrics are coherent and reminiscent of what a country song generally looks like. The model has learned to capture country music-specific vocabulary, using words like 'Tennessee', 'trucks', and 'mama'.

On closer observation, we can tell that the song even seems to follow common song structure. The inference of the song structure has been listed alongside the lyrics. As seen, the chorus seems to have a repetitive structure seen very commonly in most music, especially in the country genre. This is impressive for the level of simplicity of our model.

In conclusion, our model has performed very well for its scope and complexity. However, it will be useful to see how the model learns on a larger dataset. The model can also benefit from one or two additional layers in the future when training time is not a constraint.

VII. Project Management

All of the team members shall equally work on the tasks of data preprocessing and genre classification and shall approach the task independently in order to foster different approaches to the same problem. The descriptive analysis aspect of the project shall be divided into different analyses done by each member to maximize the outcomes of the project. The project team consists of the following members, listed alphabetically:

Atharv Subhekar - Data Handling, Pre-processing, and Evaluation, Documentation

Nidhi Sura - Lyric Generation Model

Nikhil Shah - Genre Classification Model

These are the broad areas that each team member worked on, but there was much collaboration between them during the process as well. That is to say, no component of the project was the sole contribution of a singular individual but by and large a combined effort.

VIII. Key references

[1]. Koch, K. (2022, October 27). A friendly introduction to text clustering - towards data science. Medium. https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04

[2]. Jose P. G. Mahedero, Álvaro MartÍnez, Pedro Cano, Markus Koppenberger, and Fabien Gouyon. 2005. Natural language processing of lyrics. In Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA '05). Association for Computing Machinery, New York, NY, USA, 475–478. https://doi.org/10.1145/1101149.1101255

[3]. Berger, J., & Packard, G. (2022). Using natural language processing to understand people and culture. American Psychologist, 77(4), 525–537. https://doi.org/10.1037/amp0000882

[4]. Michael Fell. Natural language processing for music information retrieval : deep analysis of lyrics structure and content. Document and Text Processing. Université Côte d'Azur, 2020. English. ⟨NNT : 2020COAZ4017⟩. ⟨tel-02587910v2⟩

[5]. Barradas, G., & Sakka, L. S. (2021). When words matter: A cross-cultural perspective on lyrics and their relationship to musical emotions. Psychology of Music, 50(2), 650–669. https://doi.org/10.1177/03057356211013390
[6]. Music Genre Classification Using LSTM: https://nbandhi.medium.com/music-genre-classification-of-lyrics-using-lstm-f5c762a1b3d
[7]. Music Genre Classification Using Lyrics: https://reinventionjournal.org/index.php/reinvention/article/view/705/659
[8]. Automatic Music Genre Classification Using Lyrics: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/reports/final_reports/report003.pdf