SMU Data Science Review

Volume 8
Number 1 *Spring 2024*

Article 1

Leveraging Transformer Models for Genre Classification

Andreea C. Craus
Southern Methodist University, ccraus@mail.smu.edu

Ben Berger
Southern Methodist University, bergerb@mail.smu.edu

Yves Hughes
Southern Methodist University, yhughes@mail.smu.edu

Hayley Horn
Southern Methodist University, hayleyhorn@gmail.com

Leveraging Transformer Models for Genre Classification

Andreea Carolina Craus[1], Ben Berger [1], Yves Hughes [1], Hayley Horn[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
{ccraus, bberger, yhughes}@smu.edu
2 Sr. Solutions Architect, Databricks,
160 Spear St 15th floor, San Francisco, CA 94105 USA
hhorn@smu.edu

Abstract. As the digital music landscape continues to expand, the need for effective methods to understand and contextualize the diverse genres of lyrical content becomes increasingly critical. This research focuses on the application of transformer models in the domain of music analysis, specifically in the task of lyric genre classification. By leveraging the advanced capabilities of transformer architectures, this project aims to capture intricate linguistic nuances within song lyrics, thereby enhancing the accuracy and efficiency of genre classification. The relevance of this project lies in its potential to contribute to the development of automated systems for music recommendation and genre-based playlist creation. Moreover, understanding the linguistic features that define distinct musical genres through transformer-based models offers valuable insights into the underlying patterns and characteristics of lyrical content. The final pre-trained transformer model chosen for the final model is called DistilBERT, which is a "distilled" version of the popular pre-trained transformer model called BERT (Biodirectional Encoder Representations from Transformers). The implications, challenges, ethical concerns, and future research are discussed and are sought to be addressed.

1 Introduction

Music streaming, further facilitated by recommendation systems, has ushered in a new era of unparalleled music access and convenience, obviating the need for physical, CDs, and lengthy file downloads. The technology revolution has stimulated music discovery, exposing listeners to new artists and diverse musical styles, and enriching musical discovery for listeners and musicians. In addition, the paradigm shift in recent years has given scholars and researchers the opportunity to delve into the intricacies of user preferences, content categorization, and recommendation systems. In an interview with CyberNews, Chris Erhardt, CEO and Co-Founder of Tundedly, an online music production platform, said "the rise of digital streaming platforms and online distribution has democratized music promotion" (Erhardt, 2023). This paper presents a thorough investigation into the efficacy of transformer model architectures and their pre-training strategies, fine-tuning techniques, and dataset variations to comprehensively evaluate their performance on genre classification. Additionally, the interpretability of these

models will be investigated to understand the features learned for effective classification.

In the music streaming world, it is evident that Spotify, with over 100 million songs available and 551 million monthly users, has become the global top-runner in music streaming services (Iqbal, 2023). Machine learning, natural language processing, and recently, artificial intelligence, are fundamental to Spotify's business. It is used not only for discovery content via recommendation and search algorithms, as well as for generating playlists, extracting audio content-rich signals for cataloging and other content-based applications, understanding voice commands, serving ads, developing business metrics and optimization algorithms, creating music with AI-assisted tools, and more (SpotifyAB, 2020). While Spotify's recommendation system is mainly powered by collaborative and content-based filtering, it is constantly evolving and the Spotify research team has published studies on the latest approaches including reinforcement learning, approximate inference, graphical models, causal inference, deep learning, time series modeling, and meta-model learning (Tomasi et al, 2023).

While some music recommendation systems, including Spotify, have started experimenting with natural language processing methods for tasks such as classification, sentiment analysis, and recommendations for lyrics, these methods have been mainly used for extracting lyrics from a song rather than extracting context or sentiment from the lyrics for generating playlists or recommendations. For instance, Genius, a platform dedicated to song lyrics and music knowledge, partnered with Apple Music for providing automated synchronized lyrics for songs. While they have attempted to incorporate Genius and lyric analysis in their recommendation system, it has been shown to recommend music that represents the more common aspects of the artist rather than the song (Barrington, 2009). In addition, genres have tended to be classified based more on the artist or album, rather than each song's characteristics. For this reason, this project analyzes each song's lyrics individually to establish more accurate and contextually aware classifications. Having one of the most extensive lyrical knowledgebases, this project will be utilizing the Genius API to extract lyrical data for genre classification and playlist creation to aid in providing more accurate classifications.

Genre classification in various media forms, such as text, audio, and video, plays a crucial role in many applications, including recommendation systems, content tagging, and information retrieval. Genre classification in the context of natural language processing (NLP) falls under the broader category of text classification. Text classification involves assigning predefined categories or labels to pieces of text based on their content. In the case of genre classification, the goal is to categorize text documents, such as song lyrics, into predefined genres or categories. The transition from rule-based systems to statistical methods marked a paradigm shift in natural language processing (NLP) and laid the groundwork for the emergence of machine learning-based approaches in various language-related tasks, such as sentiment analysis, text classification and named entity recognition (Lauriola et al, 2022). In the early stages of NLP, rule-based systems were predominant. These systems relied heavily on manually crafted linguistic rules to analyze and process text. The shift to

statistical methods emerged as a response to the limitations of rule-based systems. Statistical approaches began to gain prominence in the 1990s, leveraging probabilistic models to capture patterns in large datasets. Techniques such as Hidden Markov Models (HMMs) and Maximum Entropy Models started to replace or complement rule-based systems (Lo et al, 2023). These statistical methods demonstrated improved adaptability to diverse language patterns and were more scalable for handling larger amounts of textual data.

The history of transformer models is closely tied to the development of attention mechanisms in NLP and are known for their power in capturing complex patterns and relationships with data, which has a high potential of bringing a new dimension to music recommendation and playlist generation. Transformer models are a class of neural network architectures designed to process sequential data, such as natural language and text, that have been applied in different areas with remarkable success. Introduced by Vaswani et al. in 2017, transformers revolutionized the field of deep learning by replacing traditional recurrent and convolutional networks with a self-attention mechanism (Gillioz et al, 2020).

To address these issues, this project aims to develop a playlist recommendation system that utilizes transformer models to more accurately categorize genres with a more contextually aware approach. The idea behind using transformer models for classification tasks such as these is to incorporate these functions within an application that generates playlists based on user's specific needs. In addition, this method can be applied further than genre classification to incorporate mood, event, and other preferences. The drawbacks to using transformer models for these systems is the computational resources required makes it challenging to train larger amounts of data with larger number of features.

The idea behind using genre classification models with more contextual awareness is that it gives the user more control and leverages the power of Large Language Models (LLMs) to intelligently curate playlists and recommendations tailored to users' unique needs. While listeners can now access an ever-expanding catalog of songs and an abundance of user data, users often face information overload, making it difficult to navigate and discover music effectively. The sheer volume of options can overwhelm users, leading to decision fatigue and reduced user satisfaction. This system seeks to bridge the gap between user preferences, event characteristics, and music content, ultimately enhancing user satisfaction and engagement with music streaming services.

2 Literature Review

The literature review will serve as a foundation for understanding the state of the art in event-based playlist recommendations, guiding the development of our recommendation system and methodology, and identifying gaps and opportunities for future research.

2.1 Music Systems (MS)

Music Systems (MS) have evolved significantly over the years, reflecting the changing landscape of the music industry and technological advancements. In their early stages, in the pre-Internet era, MS primarily relied on rule-based systems that recommended music based on simple criteria like genre or artist similarity. One of the earliest examples is the "Music Genome Project", developed by Pandora in 2000, which manually annotated and categorized songs based on attributes like genre and mood (Prockup et al, 2015). These systems formed the foundation for many later developments.

The 21st century marked the transition to content-based filtering, analyzing the audio and textual attributes of music. In the early 2000s, collaborative filtering gained prominence with platforms like Last.fm. Last.fm used collaborative filtering to recommend music to users based on their listening history and the preferences of others with similar tastes (Deldjoo, 2021). In this way, users could discover new music based on what their peers were listening to. Shortly after, Pandora introduced content-based filtering, analyzing musical features to generate personalized radio stations. For instance, it analyzed a song's tempo, key, and instrumentation to create stations with similar characteristics. This marked a shift from purely collaborative to more content-driven recommendations (Schedl et al, 2018). Spotify revolutionized music streaming by integrating contextual information into recommendations. With the rise of platforms like Netflix and its recommendation algorithm, various music streaming services, including Spotify and Apple Music, began to employ advanced machine learning models to improve personalization. Natural language processing (NLP) has been used to analyze song lyrics for emotion and sentiment, enhancing recommendations based on lyrical content.

The history of music systems is a testament to the ongoing evolution of music recommendation techniques, from rule-based systems to the sophisticated, data-driven, and context-aware recommendation engines we see today. These advancements have made it easier for music enthusiasts to discover, enjoy, and share their favorite tunes in an increasingly digital and interconnected world.

2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) has played a significant role in advancing music recommendation systems and in furthering the music emotion recognition tasks (Rajendran et al, 2022). By harnessing NLP techniques, these systems can delve into the textual aspects of music data, such as lyrics, user-generated tags, and playlist descriptions. NLP aids in sentiment analysis of song lyrics, allowing recommendations to be tailored to users' emotional preferences. It also enhances the semantic understanding of music, enabling more contextually relevant suggestions. The integration of NLP in music recommendation not only refines the accuracy of recommendations but also captures the emotional and cultural context of music enhancing the user's experience (Shukla, 2017).

The lyric-based sentiment classification approach has been a popular text-based solution for song sentiment analysis. This approach adopts the vector space model

(VSM) to represent lyric text and then assigns song sentiment labels such as “light-hearted” and “heavy-hearted” (Xia et al, 2008). Language models are developed using NLP techniques, with their main purpose being to determine the likelihood of word sequences occurring in a sentence using probabilistic and statistical techniques (Singh et al, 2021). The first language models were initially based on Recurrent Neural Networks (RNNs) due to the nature of human language involving sequences of words (Lauriola et al, 2022).

2.3 Attention Mechanisms

The idea of attention mechanisms was introduced by Bahdanau et al. in the paper "Neural Machine Translation by Jointly Learning to Align and Translate" in 2014. Attention mechanisms allowed models to selectively focus on different parts of the input sequence when generating an output, providing a solution to the limitations of traditional sequence-to-sequence models (Niu et al, 2021).

There are two types of attention mechanisms commonly used in transformer models, as well as artificial networks and neural network models, particularly in the context of deep learning and NLP. These mechanisms are inspired by the way attention works in the human brain. On one hand, there is bottom-up attention, which is driven by the saliency of input features and focuses on the most relevant and prominent features in the input data, and on the other hand, there is top-down attention, which is guided by higher-level knowledge or context and involves focusing on specific parts of the input based on the task at hand or prior information (Niu et al, 2021). There are four different criteria that can be used to categorize attention mechanisms, along with the types of attention within each criterion (Niu et al, 2021), which can be seen in Table 1.

Table 1. Four different criteria used to categorize attention mechanisms and the types of attention within each criteria.

Criteria	Type
The Softness of Attention	Soft or Hard, Global or Local
Forms of Input Features	Item-wise or Location-wise
Input Representations	Distinctive, Self, Co-Attention, Hierarchical
Output Representations	Single-Output, Multi-Head, Multi-Dimensional

In the context of this analysis, which is classifying lyrics for songs, or *documents*, this falls under the document classification category which uses a soft attention mechanism, item-wise input features, hierarchical input representations, and single-output representations (Niu et al, 2021). In document classification, the document, or the lyrics in this case, is assigned a label based on the category, with the goal being to categorize the text automatically into the predefined categories based on its content (Patwardhan, 2023).

2.4 Deep Learning

Deep learning is a broader concept that encompasses a class of machine learning techniques that utilize neural networks with multiple layers called deep neural networks, which are capable of learning hierarchical representations from data and automatically extract features at different levels of abstraction (LeCun et al, 2015). The term “deep” refers to the depth of a neural network which indicates the presence of multiple layers between the input and output layers. Deep learning have a wide range of architectures, including Feedforward Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory, and Transformers (Gupta et al, 2020). These models are fundamentally about learning meaningful representations of data, which capture hierarchical and contextual information, enabling the model to make accurate predictions, as well as generate relevant output.

2.5 Transformer Models

The transformer architecture was introduced by Vaswani et al. in the paper "Attention is All You Need" in 2017. The transformer model eliminated the need for recurrent or convolutional layers, relying solely on self-attention mechanisms. This architecture demonstrated superior performance in machine translation tasks and quickly became a breakthrough in the field of NLP.

In 2018, the Bidirectional Encoder Representations from Transformers (BERT) model was introduced. BERT significantly advanced the state-of-the-art in various NLP tasks by pre-training a transformer-based model on large amounts of unlabeled text (Devlin et al, 2018). BERT demonstrated the effectiveness of bidirectional context understanding and contextualized word embeddings, paving the way for further transformer-based models. OpenAI introduced the Generative Pre-trained Transformer 2 (GPT-2) model in 2019. GPT-2 demonstrated impressive language generation capabilities and the ability to perform a wide range of NLP tasks. GPT-2 was known for its large number of parameters, contributing to its exceptional performance.

Transformer models have made significant contributions to NLP tasks including sentiment analysis, named entity recognition, part-of-speech tagging, text summarization, and machine translation (Patwardhan et al, 2023). In addition, they have excelled in question answering tasks by understanding and extracting relevant information from context. For example, BERT has been used by various QA systems to provide accurate and contextually aware responses. Transformer models like GPT-3 have been renowned especially for their text generation capabilities. This is one application that has been researched for lyrics generation as well.

Pre-trained transformer models are deep learning models that have been trained on large amounts of data and then fine-tuned for specific tasks, leveraging a very large dataset to allow the model to learn general features and representations of data (Gong et al, 2020). These pre-trained transformer models follow the typical transformer architecture which implements self-attention layers, feedforward layers, and

normalization layers (Singh et al, 2021). The pre-training data set usually contains unlabeled or weakly labeled data.

DistilBERT Pre-Trained Transformer Model. The pre-trained transformer model used for this analysis is called DistilBERT, short for "Distilled BERT", which is a more efficient and lighter version of the BERT model (Sanh et al, 2019). It is designed to retain much of the performance of BERT while significantly reducing the number of parameters, making it more computationally efficient and better suited for resource-constrained environments. This pre-trained model is created through a process known as distillation which involves training a smaller model (DistilBERT) to mimic the behavior of a larger, pre-trained model (BERT). In the case of DistilBERT, the size of the smaller model is reduced by 40% compared to the BERT model, "while retaining 97% of its language learning capabilities and being 60% faster" (Sanh et al, 2019). During this distillation process, the smaller model learned to replicate knowledge encoded in the attention mechanisms and hidden layers of the larger model. The parameter reduction in the smaller model is achieved through various architectural simplifications including fewer layers in the transformer, fewer attention heads, and a reduced hidden state dimension in comparison to BERT. However, despite these significant reductions, DistilBERT is still able to capture complex linguistic patterns. When trained on an IMDB dataset for a sentiment classification task, DistilBERT was only 0.6% behind BERT in test accuracy (Sanh et al, 2020).

3 Methodology

3.1 Data

The dataset was retrieved from Kaggle as two CSV files, one with information on the artists including the genre for each artists, and another file containing the lyrics along with the language associated with the song. These two files were merged based on the artist, resulting in 267,259 songs and 56 different genres, with some songs having multiple genres associated with them. While including more data and additional genres to the analysis would definitely improve the results, for the purpose of this analysis and the available system resources, this project will focus on only on four genres: Rock, Indie, Country, and Hip-Hop/Rock. The dataset was filtered on these four genres, as well as filtering on only English songs. In addition, two duplicate entries for the same song were found which were removed, resulting in 17,160 song lyrics in the dataset for analysis. There were no missing values or lyrics present in the data following this filtering process.

Data Preprocessing. This step plays a crucial role in enhancing the quality and relevance of the insights derived from textual data, especially in the context of analyzing lyric data. The process of transforming raw lyric text through various techniques serves to address several key challenges inherent in natural language processing and aids in capturing the nuances of different contexts of the text.

The function created to clean the text of the lyrics was tailored for the specific characteristics of lyric data, performs essential tasks to preprocess and standardize the text, which are shown in Figure 1. First, the text is converted to all lowercase to ensure uniformity. It then removes any HTML tags present in the lyrics, addressing potential artifacts gathered during web-scraping and the data collection process. The numerical values within the lyrics are replaced with a placeholder string "NUM", in order to abstract them and allow the model to include contextual context rather than focusing on the specific numerical values. While in some cases removing the numerical values completely does not affect the model, the contextual information that numbers may be able to provide is valuable in the context of this lyrical classification task. In addition, the question marks and exclamation points were also replaced with a placeholder string "SYMB", simplifying punctuation variations but retaining the emotion or context of what the question marks and exclamation points could signify. Finally, the additional punctuation and special characters are removed to enhance the cleanliness of the text data. These steps collectively contribute to a cleaner and more standardized representation of the lyric content.

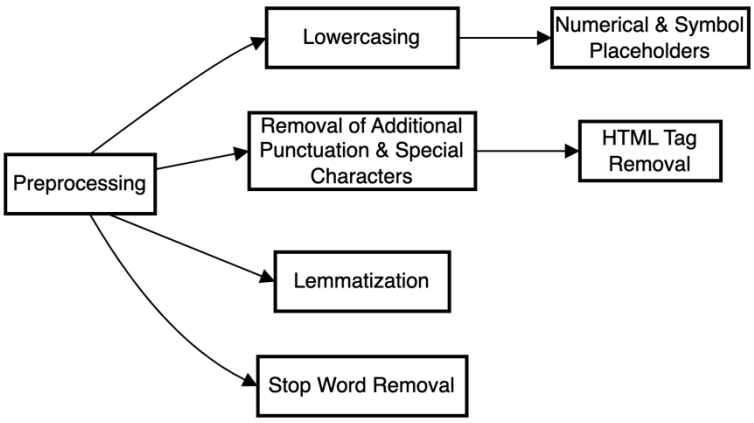

Fig. 1. The steps employed in the data preprocessing phase starting with (1) Lowercasing and numerical and symbolical placeholders, (2) Removal of additional punctuation and special characters as well as HTML tags, (3) Lemmatization, the process of converting words to their root form, and finally (4) Removal of stop words, which are common words that give little value to the meaning of the context of text.

The next step in the preprocessing is performing lemmatization on the lyric text, which is the linguistic process of reducing words to their base or root form, called the "lemma", which represents the canonical, or dictionary, form of a word. This process is important because it helps normalize the words by reducing them to their base forms, reducing the dimensionality of the vocabulary, which ensures that the different forms of a word, such as plurals or verb conjugations, are treated as the same token in the model. This allows the model to generalize better across variations of a word and

ensures the model recognizes the semantic equivalence of words in different forms. An additional benefit to performing lemmatization is that it can contribute to feature reduction, so the model has fewer unique tokens to process which makes it more computationally efficient, as well as reduces the risk of overfitting.

Next in the preprocessing is the removal of stop words, which is filtering out and removing common words that are considered to be of little value in representing the content and meaning of the text. These words are frequently occurring terms like “and”, “the”, “is”, and “in” that do not carry significant semantic meaning. This process is important because by removing the less meaningful words from the lyrics, the focus shifts to the more meaningful terms, which reduces the noise in the data and allows the model to focus on the content-rich words. In addition, since these words and usually common across different genres and contexts, including them in the analysis has potential to lead to overfitting, making the model too specific to the training data and performing poorly on new, unseen data. Removing stop words can enhance the ability for the model to generalize to diverse texts. In addition, as these words are extremely common, their presence in the text can significantly increase the dimensionality of data, therefore removing them helps reduce the feature space, making computations more efficient and speeding up the time it takes to train the model.

For each of the four genres, the distribution of the number of words in each of the individual lyrics following each step of the preprocessing is shown in Figure 2. The initial lyrics are shown in blue, although mostly overlapped by the distribution of the initial cleaned lyrics (yellow) and the lemmatized lyrics (green), this shows how each step of the process is narrowing down the number of words each of the lyrics has and automatically eliminating many outliers, which are important tasks for preparing text for use in transformer models. There is a clear distinction when it comes to the distributions following the removal of stop words, which is shown in red, essentially halving the number of words in each of the lyrics.

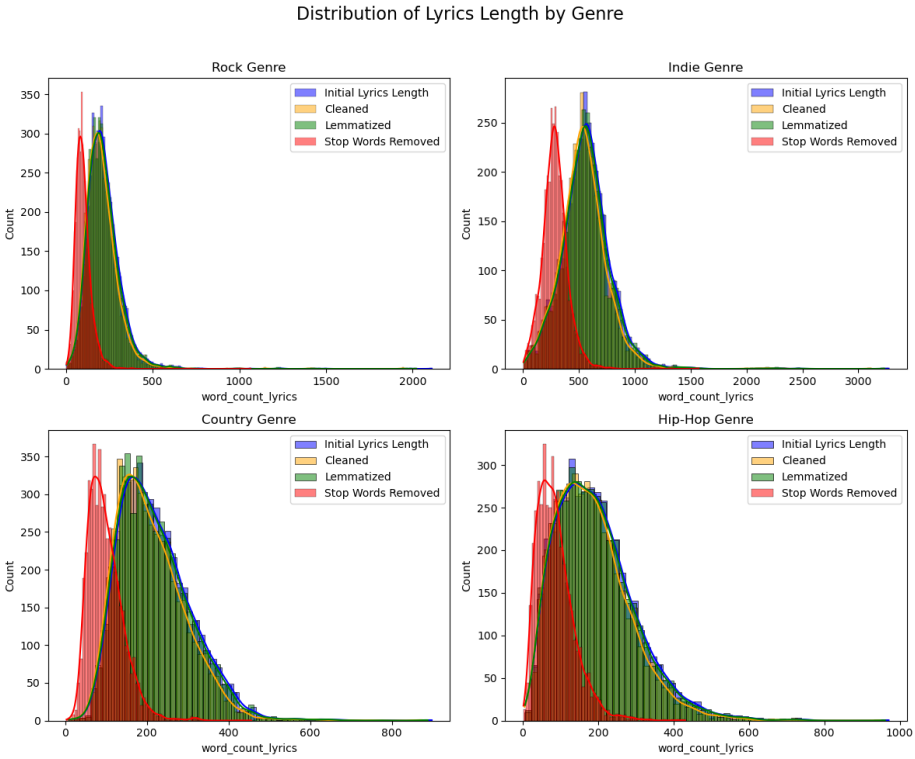

Fig. 2. Histograms for the distribution of length of the individual lyrics for different steps of the preprocessing phase showing the original length (*blue*), the length following the general cleaning of the text including removal of punctuation, symbols, and HTML tags (*yellow*), the length following the lemmatization step (*green*), and the final length after the removal of stop words (*red*).

In the context of lyrical analysis, another important step in preparing the data is Part-of-Speech (POS) tagging. This process involves assigning a grammatical category (such as noun, verb, adjective, etc.) to each word in the text to analyze and understand the syntactic structure of sentences, identifying the role of each word in the context of the sentence. This semantic understanding is crucial for discerning the nuances of different genres, as the choice and arrangement of words contribute significantly to the overall meaning of a lyric. Transformer models are designed to capture contextual information by considering the entire context of a word within a sentence and POS tagging further enhances this contextual understanding, enabling the model to better grasp the syntactic nuances that contribute to genre-specific writing styles.

Exploratory Data Analysis (EDA). In Figure 3, the word clouds of for the most prevalent words in the "Country", "Indie", "Rock" and "Hip-Hop" rap genre are shown.

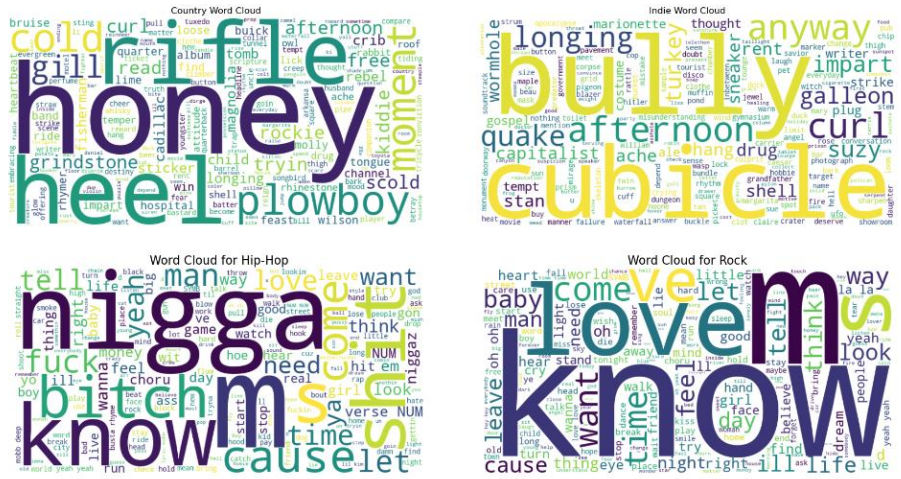

Fig. 3. Words clouds for each four genres in the analysis: Country (*top-left*), Indie (*top-right*), Hip-Hop (*bottom-left*), and Rock (*bottom-right*).

[1] It's crucial to acknowledge that the inclusion of potentially offensive terms is not intended to endorse or condone their usage but rather to faithfully represent the linguistic landscape present in the dataset. Moreover, efforts have been made to handle these terms with sensitivity and discretion, ensuring that their inclusion serves the academic purpose of the study without causing harm or offense.

These word clouds display the words in varying sized based on their frequency in the dataset, showing certain patterns in the lyrics of certain genres and understanding overall themes or topics. In addition, these give important insights into outliers or unusual terms to determine if additional special attention or cleaning is necessary and understanding the potential noise or anomalies in the dataset.

For conducting a genre classification analysis, it's essential to maintain a comprehensive and unbiased representation of the textual data. While this task necessitates the exploration of various linguistic elements, including potentially offensive terms, it is important to clarify the rationale behind their inclusion. In the figure above, certain explicit terms, which may be deemed offensive or sensitive, have been included in the word cloud analysis. The decision to incorporate such terms stems from their significant presence within the dataset and potential relevance to the genre classification analysis. Omitting these terms could lead to a skewed understanding of the underlying themes and language patterns within the data, thereby compromising the integrity and accuracy of the analysis.

3.2 Design and Procedure

The general design and procedure for the DistilBERT model can be seen in Figure 4, beginning with data preparation and ending in hyperparameter tuning.

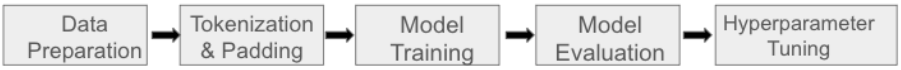

Fig. 4. The general design and procedure for the modeling process: Data Preparation, Tokenization and Padding, Model Training, Model Evaluation, and Hyperparameter Tuning.

The first step of the of the procedure is the data preparation in which the dataset of lyrics with corresponding genre labels are collected and pre-processed. The text representation process is crucial when using transformer models for classification and NLP tasks. Each lyric is converted into a numerical format that the model can understand. This process involves tokenization to break down the lyrics into individual words or subwords and then uses an embedding layer to map the tokens to continuous vector representations. The dataset is then split into training, validation, and testing sets, which is essential for maintaining a fair distribution of the target variable for the model to generalize well to new data and provide more reliable and robust predictions in real-world scenarios. The training set is used to train the model and the testing set is used to assess how the model generalizes to new, unseen data. This separation helps to avoid overfitting.

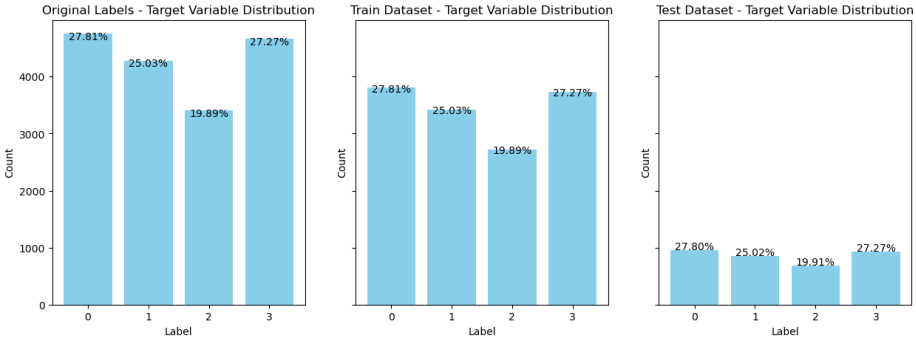

Fig. 5. Bar charts showing the target variable distribution for the original labels, the training dataset, and the test dataset. The numerical labels correspond to the genres Country (0), Indie (1), Hip-Hop (3), and Rock (4).

As can be seen in Figure 5, the target variable in the entire dataset, as well as the training and testing sets, is equally distributed across all labels. Ensuring a fair distribution of the target variable across the training and testing sets is crucial for model training and evaluation to avoid bias and ensure the model is not only good at predicting certain classes and performing poorly on others.

Transformer Architecture. The general transformer architecture has become a foundational structure for many NLP tasks and is known for its ability to capture long-range dependencies in sequences efficiently (Vaswani et al, 2018). For sequence-to-sequence tasks, such as in this analysis, the transformer model has an encoder-decoder architecture, although it is also applicable for tasks where only the encoder or decoder is required, such as language modeling, in this case. This part of the architecture pertains to the training method employed to generate embeddings from input tokens (Irie et al, 2019).

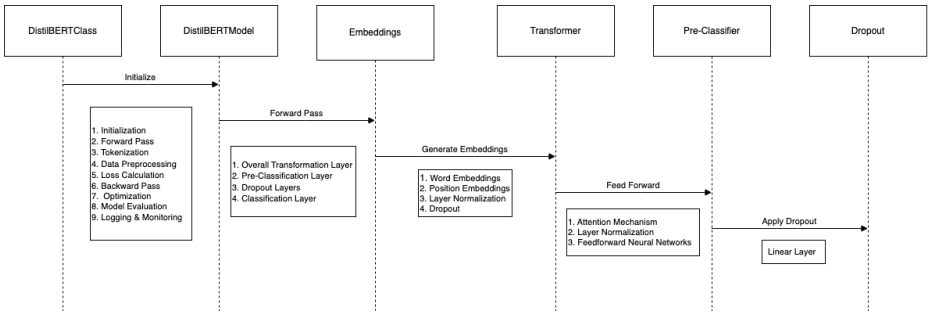

Fig. 6. The DistilBERT architecture employed in the modeling phase (*left to right*): DistilBERT Class, DistilBERT Model, Embeddings, Transformer, Pre-Classifier, Dropout.

For the DistilBERT model used in this analysis, the general architecture can be seen in Figure 6. The DistilBERT class represents the architecture of the DistilBERT model

wrapped within a custom PyTorch class called "DistilBertClass". This class is defined to take in the data as input and generate tokenized output for the DistilBERT model using the DistilBERT tokenizer for the lyrics and creates 2 datasets, for training and validation. The DataLoader is then used to create the training and validation data loaders in order to load the data to the neural network in a defined manner, which is necessary due to the size of the data not being able to be loaded to the memory at once. The embedding layer contains word embeddings, which map each word in the input vocabulary to a vector, and position embeddings, which represent the positional information of words in the input sequence. Layer normalization is then applied to the embeddings as well as the dropout with a probability of 0.1.

Loss Function. A crucial component in deep learning and in implementing transformer models is the loss function, which is used to quantify the difference between the predicted output of a model and the actual target values during training. In the context of this project and text classification problems using transformer models, the loss function commonly used is the cross-entropy loss, which measures the dissimilarity between the predicted probability distribution and the actual distribution (Singh et al, 2021). The cross-entropy loss formula (1) is given by the equation below, where:

- i represents each genre category in the classification task
- y_i is a binary value indicating whether the correct genre label is assigned to the input text, or lyrics.
- pi represents the predicted probability assigned by the transformer model to the input text, or lyrics, belonging to the genre category i

$$\textit{Cross-Entropy Loss} = -\sum_i y_i \cdot \log(p_i) . \quad (1)$$

In fact, the main goal during training for transformer models is minimizing the loss function, which is essentially measuring the model's performance and aids in adjusting the model's parameters, called weights, through the process of backpropagation.

3.3 Modeling and Evaluation Metrics

This research employed a transformer model designed to accurately classify music genres from lyrics of songs. The model was selected based on its ability to handle complex patterns found in music, with a focus on achieving high accuracy and robustness against overfitting. The final model leveraged a combination of feature engineering techniques to capture the nuances of musical genres.

Model Parameters. The parameter configuration process for the DistilBERT model is crucial as these settings directly influence the behavior of the model during training. There are various hyperparameters that can be specifically set and tuned for the specific task at hand. In addition, the choice of settings for these hyperparameters can significantly affect the computational resources and time training would take. While

increasing the model size, such as the hidden size or number of layers increases the capacity of the model, it also increases computational requirements. The dropout and model depth parameters influence the model's ability to generalize on unknown data as well as prevent overfitting (Vaswani et al, 2017). Another important parameter is the number and size of attention heads which impact the model's ability to capture different types of information, dependencies, and attention patterns. In addition, there are parameters like max position embeddings, which influence the model's ability to handle longer sequences, and parameters for layer normalization and initializer range, which affect the stability and convergence speed while training the model. A few of the important hyperparameters that need to be set before training are described below:

- `BatchSize:` Controls the number of samples from the training data that is used on each training step. For each training step, the predictions are compared with the actual values, the error is calculated, and the hyperparameters are tuned (Barbon et al, 2022).
- `Epochs:` Contains of at least one batch and controls the number of times the training data will pass through the model during the training process
    - The larger the epoch value, the higher the chance the model will undergo overfitting and not generalize to unseen data. The lower the epoch value, the higher the change the model will undergo underfitting, meaning the model has not learned enough from the training data.
- `Learning Rate:` This parameter is essential for regulating the adjustment magnitude of the model weights during training iteration.

Evaluation Metrics. To assess the performance of the model and reduce the risk of overfitting, cross-validation was used to divide the dataset into multiple subsets. The model was then trained on different combinations of these subsets and the performance was evaluated across various splits. Cross-validation provides a more robust estimate of the model's performance compared to a single train-test split. By averaging the performance metrics across multiple folds, it offers a more reliable indication of how well the model is likely to generalize to unseen data. As a key metric, the validation loss quantifies the model's error on the validation set. A final average validation loss of 0.808 was achieved, indicating the model's effectiveness in minimizing errors when predicting musical genres. This metric is crucial for fine-tuning model parameters and selecting the most efficient model architecture.

4 Results

The final model was only trained on one epoch due to the computational intensity of the training process, with a final average validation loss of 0.808 and a final average accuracy of 62.41%, which is adequate. In Figure 7, a confusion matrix for the model prediction is shown which shows the number of correctly predicted labels for each genre as well as the number of incorrectly predicted labels for each genre.

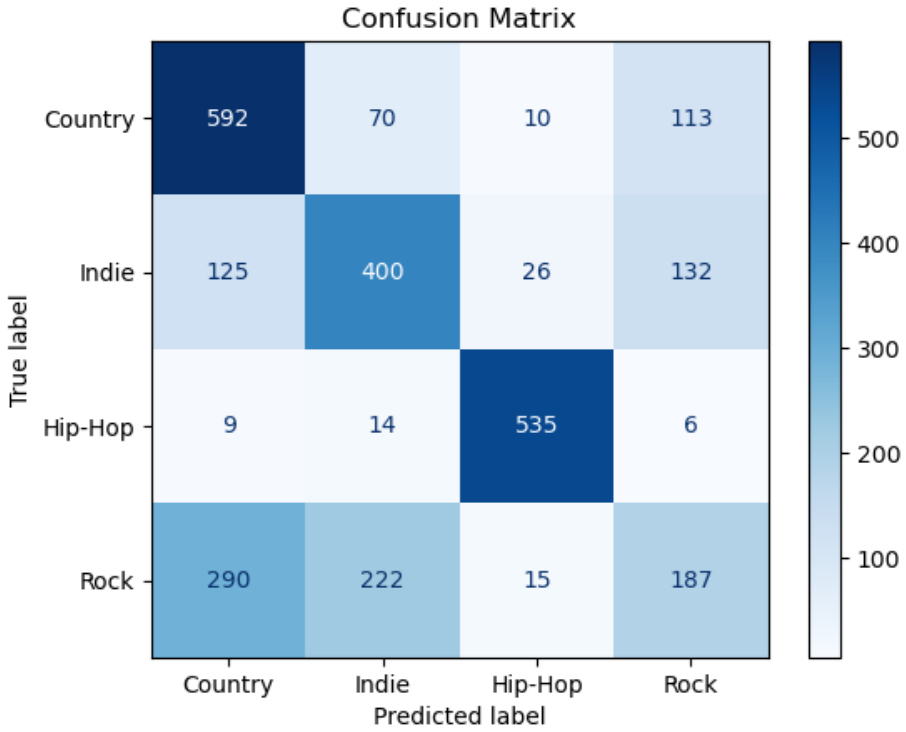

Fig. 7. Confusion matrix shows the correctly and the incorrectly classified labels for each genre which provides a comprehensive overview of the performance of the genre classification across the various genres.

In the confusion matrix depicted in Figure 7, each row in the matrix represents the actual genres, while each column represents the predicted genres. The cells of the matrix contain the frequency of instances classified into each category:

- True Positive (TP): Instances correctly classified to a specific genre.
- False Positive (FP): Instances incorrectly classified to a specific genre.
- True Negative (TN): Instances correctly classified as not a specific genre.
- False Negative (FN): Instances incorrectly classified as not a specific genre.

5 Discussion

5.1 Limitations

While the research uncovered in this project is valuable, it comes with certain limitations due to data access limits and computational resources. First and foremost, if

additional data were collected, the accuracy and efficacy of the model would improve even further. In addition, the researchers had access to limited infrastructure to handle large-scale datasets or user bases to scale as well limited computational resources that some of the complex recommendation algorithms require.

The project initially sought to classify and recommend songs based on user event preferences, however, there are no pre-made datasets with this sort of information, nor is this information easily scrapped using an API, and none of the music platforms or websites provide consistent event-labeling in order to have the ability to do cross-validation on a test set. This would require manually labeling a dataset, which would take more time and resources than available. While a general sentiment analysis can be done fairly accurately on the lyrics, the outcomes are usually just positive or negative sentiment.

An additional significant limitation in using transformer models is the computational intensity and resources they require. Transformers, especially larger pre-trained models, require substantial computational power and memory, making them resource-intensive and less accessible for small research teams and limited computational resources and budget. In addition, the extensive number of parameters available in these models can lead to long training times and therefore difficulties in fine-tuning for more specific tasks.

5.2 Implications and Challenges

Many challenges are presented when working with large amounts of data and limited resources. Obtaining good lyric data for a detailed analysis poses several challenges, which can impact the quality and robustness of the results. Access to comprehensive and high-quality datasets is limited due to copyright and licensing issues. Many available datasets are user-generated and created by using APIs, such as the LyricsGenius API, or scraped from websites such as LyricsGenius.com. Lyric websites often leverage user-generated content, leading to variations in quality, accuracy, and completeness of the individual lyrics. Inaccuracies, missing lines, and misinterpretations can be common in scaped data or user-generated content. In addition, websites frequently update their structure, and the layout is inconsistent.

An additional implication when analyzing lyrics is due to multilingual content and language variability as some lyrics, even when filtering by country, are often in various language, as well as lyrics that may be in multiple languages, each with its own unique challenges in terms of linguistic nuances and character encoding. Handling multilingual content requires additional considerations that this project does not delve into, but that would likely improve the accuracy of the model and analysis. Once the data was obtained, intensive preprocessing and cleaning was necessary before the data could be used in the model which is time intensive. In addition, working with such large amounts of data makes it difficult to comprehensively analyze if the lyrics are accurate, complete, and consistent throughout the dataset.

5.3 Ethical Concerns

Our research has brought forth several ethical considerations that warrant careful attention. As we leverage the deployment of transformer models in music genre classification and recommendation systems, we must consider how this could be used in practice for unethical purposes driven by profit or power. For example, could a playlist be created that could knowingly drive someone to purchase a product?

Privacy and Data Security. During this research, we prioritized privacy and security by utilizing publicly available datasets from Kaggle for genre classification. These datasets did not contain any personal user data, mitigating privacy concerns related to individual listening habits or preferences. By relying on aggregate and anonymized data, we adhered to ethical research practices that respect user privacy.

Algorithmic Bias and Fairness. Our research deliberately employed a dataset sourced from Kaggle that represents a broad spectrum of music genres. This diverse dataset was chosen to minimize bias towards a particular genre, ensuring a more balanced representation in our genre classification model. Throughout the research process, we employed techniques such as stratified sampling to maintain genre diversity in training and testing sets, aiming to detect and mitigate any inadvertent biases early on. Additionally, our analysis included rigorous validation procedures to assess the model's performance across different genres, identifying any disparities that could indicate underlying biases.

Cultural Diversity and Representation. Our approach to ensuring cultural diversity and representation within this project involved the deliberate selection of a dataset from Kaggle that encompasses a wide array of genres, including Rock, Indie, Country, and Hip-Hop/Rap, among others. This diverse dataset allowed us to train the DistilBERT model on a broad spectrum of lyrical content, aiming to capture the unique linguistic features that characterize different musical traditions. By incorporating a varied dataset, the research not only aimed at high accuracy in genre classification but also sought to reflect the rich diversity of global music genres. This methodology underscores our commitment to representing a wide range of musical expressions, thereby reducing the risk of cultural bias that often plagues algorithmic recommendations.

User Autonomy and Personalization. The core of our project's methodology focused on leveraging the DistilBERT transformer model for music genre classification, aiming to understand and categorize musical content with high precision. A key aspect of this research was the potential application of these insights to enhance user personalization in music recommendation systems. While our project primarily dealt with genre classification without direct user interaction, the implications of our findings for personalization are significant. By accurately identifying genres, we can enable more refined personalization, catering to the nuanced preferences of diverse user bases.

5.4 Future Research

The potential areas for future research and improvement in using transformer and compound models in music streaming platforms is vast and emerging technologies will continue to enhance these types of systems.

This project only explored the classification of four genres, the first future area of research would be adding additional genres and subgenres to the classification task. Another area of further research would also be to perform more extensive cleaning and preprocessing of the lyrics as that would also further improve the accuracy of the predictions. In addition, a multi-label classification for the genres could provide interesting insights and more accurate results. With this type of transformer model, such as DistilBERT, additional features can also be studied such as audio features, tags, descriptions, and many others, to further expand the knowledge of the model. These models can be extended to find patterns within songs based on mood, event, and other preferences as well.

While initially setting out to create a recommendation system, it is difficult to access user data for privacy reasons, simplifying this analysis to a classification task, however, with the necessary access and computational resources, the future research consists of creating a playlist recommendation system that users have more control of. Another intriguing application to further research is a song library organizational tool that generates playlists from a user's library based on settings the user can tweak to their liking.

Future research is likely to explore innovative approaches to enhance personalization, context-awareness, and user engagement. This includes deep learning architectures, such as graph neural networks, to capture intricate patterns in music data and user behavior. Another area for future research is multimodal recommendations which could combine audio features, user interactions, and visual content, such as album covers or music videos, to create more holistic and expressive music recommendations.

Incorporating user input into current technology can significantly enhance personalized musical experiences. Spotify's DJ AI tool, a feature within the app that crafts dynamic playlists based on user preferences, currently lacks the capability for user input. One notable limitation is the inability to exclude specific genres from the DJ's selection process. For example, if a user frequently listens to sleep music to aid in falling asleep, the DJ may misinterpret this preference, assuming the user enjoys sleep music throughout the day. Consequently, the generated playlists may feature an abundance of sleep music, which may not align with the user's actual listening habits. Allowing users to provide feedback and exclude certain genres would refine the AI's understanding of individual preferences, leading to more accurate and satisfying playlists tailored to each user's unique tastes and habits.

6 Conclusion

This research embarked on an exploration of leveraging transformer models, specifically DistilBERT, for the nuanced task of music genre classification. This journey was guided by a desire to harness the sophisticated capabilities of transformer architectures to capture the rich linguistic textures of song lyrics, aiming to contribute meaningfully to the fields of music recommendation and playlist creation. Through meticulous methodology, encompassing data collection, preprocessing, and model training, it was demonstrated that transformer models could indeed offer a significant advantage in understanding and classifying musical genres based on lyrical content.

Reflecting on the ethical considerations, this research underscored the importance of navigating the challenges associated with privacy and data security, algorithmic bias and fairness, cultural diversity and representation, and user autonomy and personalization. It should be acknowledged that the application of AI in music recommendation systems carries profound implications, not just in enhancing user experience through personalization, but also in ensuring that such innovations are pursued responsibly, ethically, and inclusively.

Looking ahead, the potential for future research in this area is vast. While this study focused on genre classification, the application of transformer models in music recommendation systems opens avenues for more personalized, context-aware, and culturally diverse music discovery experiences. The exploration of multimodal data, deeper engagement with ethical AI practices, and the continual quest for balance between personalization and user autonomy present exciting opportunities for advancing the state of the art in music recommendation technologies.

In conclusion, this research contributes to the ongoing dialogue on the role of AI in music analysis and recommendation, highlighting both the technological potential and the ethical responsibilities inherent in this domain. As technology continuously evolves, it is our hope that this work will inspire further exploration, innovation, and thoughtful consideration of how AI can be leveraged to enrich musical landscapes while honoring the diverse tapestry of global musical heritage and respecting the rights and preferences of users worldwide.

Acknowledgments. Hayley Horn, Capstone Advisor

References

1. Barbon, R. S., & Akabane, A. T. (2022). Towards transfer learning techniques—Bert, Distilbert, Bertimbau, and Distilbertimbau for automatic text classification from different languages: A case study. Sensors, 22(21), 8184. https://doi.org/10.3390/s22218184
2. Barrington, L., Oda, R., & Lanckriet, G. R. (2009, October). Smarter than Genius? Human Evaluation of Music Recommender Systems. In ISMIR (Vol. 9, pp. 357-362). https://archives.ismir.net/ismir2009/paper/000014.pdf

3. Deldjoo, Y., Schedl, M., & Knees, P. (2021). Content-driven music recommendation: Evolution, state of the art, and challenges. arXiv. doi:https://doi.org/10.48550/arXiv.2107.11803
4. Erhardt, C. (2023, September 21). Chris Erhardt, tunedly: "The rise of digital streaming platforms and ... CyberNews. https://cybernews.com/security/chris-erhardt-tunedly-the-rise-of-digital-streaming-platforms-and-online-music-distribution-has-democratized-music-promotion/
5. Gillioz, A., Casas, J., Mugellini, E., & Khaled, O. A. (2020). Overview of the Transformer-based models for NLP tasks. *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*, 179–183. https://doi.org/10.15439/2020f20
6. Iqbal, M. (2023, August 2). Spotify revenue and Usage Statistics (2023). Business of Apps. https://www.businessofapps.com/data/spotify-statistics/
7. Irie, K., Zeyer, A., Schlüter, R., & Ney, H. (2019). Language modeling with Deep Transformers. Interspeech 2019. https://doi.org/10.21437/interspeech.2019-2225
8. Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and Tools. Neurocomputing, 470, 443–456. https://doi.org/10.1016/j.neucom.2021.05.103
9. Lo, A. W., & Singh, M. (2023). From Eliza to ChatGPT: The evolution of natural language processing and financial applications. The Journal of Portfolio Management, 49(7), 201–235. https://doi.org/10.3905/jpm.2023.1.512
10. Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of Deep Learning. Neurocomputing, 452, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091
11. Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the real world: A survey on NLP applications. Information, 14(4), 242–248. https://doi.org/10.3390/info14040242
12. Prockup, M., Ehmann, A. F., Gouyon, F., Schmidt, E. M., & Kim, Y. E. (2015). Modeling Musical Rhythm at scale with the Music Genome Project. 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 1–5. https://doi.org/10.1109/waspaa.2015.7336891
13. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv. https://doi.org/arXiv:1910.01108v4
14. Schedl, M., Zamani, H., Chen, CW. et al. Current challenges and visions in music recommender systems research. Int J Multimed Info Retr 7, 95–116. doi:https://doi.org/10.1007/s13735-018-0154-2
15. Shukla, S., Khanna, P., & Agrawal, K. K. (2017). Review on sentiment analysis on music. *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, 777–780. doi:https://doi.org/10.1109/ictus.2017.8286111
16. Singh, S. (2022). Music recommendation system using content and collaborative filtering methods. INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 06(05). doi:https://doi.org/10.55041/ijsrem12733
17. SpotifyAB. (2020). Machine learning. Spotify Research. https://research.atspotify.com/machine-learning/
18. Tomasi, F., Cauteruccio, J., Kanoria, S., Ciosek, K., Rinaldi, M., & Dai, Z. (2023). Automatic Music Playlist Generation via simulation-based reinforcement learning. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 4948–4957. doi:https://doi.org/10.1145/3580305.3599777

19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. doi: arXiv:2212.01020
20. Xia, Y., Wang, L., & Wong, K.-F. (2008). Sentiment vector space model for lyric-based song sentiment classification. International Journal of Computer Processing of Languages, 21(04), 309–330. doi:https://doi.org/10.1142/s1793840608001950