Music Recommender System using Natural Language Processing and Audio Feature Analysis

MSc Research Project
M.Sc. in Data Analytics

Kunal Chauhan
Student ID: 21139245

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Kunal Chauhan		
Student ID:	21139245		
Programme:	M.Sc. in Data Analytics	**Year:**	2022-23
Module:	M.Sc. Research Project		
Supervisor:	Qurrat Ul Ain		
Submission Due Date:	15 December 2022		
Project Title:	Music Recommender System using NLP and Audio Features Analysis		
Word Count:	5635	**Page Count**	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Kunal Chauhan

Date: 14 December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Music Recommendation System Using Natural Language Processing and Audio Feature Analysis

Kunal Chauhan
ID:21139245
Research Project
MSCDAD JAN22 B

Dec 2022

Abstract

In this research an attempt is made to build a Music Recommendation system which includes both the main features of music i.e. lyrics and audio. The lyrics will be analysed using Natural Language processing technique and sentiment analysis will be done for the recommendation. Also simultaneously audio feature extraction of timbre audio features will be done using libROSA and classification model will be built using machine learning and deep learning models for the genre classification and prediction. CNN is used for this study as the main classification technique and has yielded the best results.

1 Introduction

1.1 Background

Personalisation is a fairly new concept which has taken over the world specially in the entertainment industry. This concept has gained popularity in today's era because of the fast-evolving technology and the awareness in people nowadays about their choices. This awareness has also increased because of the availability of the content is in abundance and people have a very fast life. Back in earlier days people were limited to the cable TVs and their choices had to be aligned with cable provider as they had no choice but to watch whatever was being broadcasted to their screens. But the scenarios have changed now and we now have something called OTT Platforms which give people a choice as to watch whatever they want to and whenever they want to. Now the dependencies have reduced and a sort of freedom has increased. The motive behind narrating this was that this sort of freedom has been achieved in the Music industry as well. For music we have numerous free sources as well as paid platforms like YouTube, Spotify, and many more to name a few. They give the users content in vast abundance to choose from according to their mood, their taste and their time. The basic function of a recommender system as the name suggests provides recommendations of the similar sort of things that the user is in a habit of exposing itself. There have been numerous music platforms designed using the recommendation system algorithms which are popularly clubbed and known as Music Recommender Systems (MRS).

1.2 Motivation

Music Recommender System actually works on a very simple algorithm of similarity. Now, this can be based on two things and includes two prominent features of a song which are namely the lyrics and audio-features. As a song is mainly composed by lyrics and sound, research will focus on both one by one.
The motive behind the research study is to build a recommendation system using Natural Language Processing tools to find and analyse the sentiments of the lyrics. This is of utmost importance because the soul of a song are the lyrics used in it and more importantly the context conveyed in the song by the lyrics. Without this step the audio features extraction alone won't be of much help because in life as well you need words to express your sentiments and understand what the person is saying. The second part is the extraction and analysis of the audio features. In this part the audio signals will be turned in to images and a pictorial representation of the sound waves will be done using Fourier Transform and these images will be pushed in our CNN and other machine learning models. The combination of the analysis of the sentiments through lyrics and the extraction of audio features will be used make recommendations for the user in the MRS. Hence sentiment analysis through lyrics will lay the foundation for the above mentioned MRS. Basis of these two steps a library of pre-existing genres will be built and predictions will be done. Genre classification and prediction is done by using machine and deep learning techniques. The research presented has profoundly not used any methods like collaborative filtering with a very strong reason that actually the functioning of collaborative filtering is that it will take recommendations on the choice via other user in lay man terms. This, in a more technical language, means that Users X and Y will rate or act on other items similarly if they rate n items similarly or exhibit similar behaviour. Instead of determining how similar two products are, a group of ”nearest neighbour” users for each user are identified

based on how strongly their past evaluations correlate. As a result, scores for the things that cannot be seen are predicted using a mixture of scores from the closest neighbours. So, if we look closely and understand the meaning of collaborative filtering or matrix factorization which is again a class of collaborative filtering is that to begin with, the essential requirement will be a user whose data will be used to compare and find the contrast. But when it comes the study in hand to be evaluated, the essence of it lying in finding an alternative to that and working our way around the genres and creating a pre-existing history of the genres which is in turn used to actually make the recommendations. By this way we don't have to rely on a user to create its history first so that the recommender system can make the recommendations.

1.3 Research Question

How can Music Recommender System be built using Natural Language Processing and Audio Feature Analysis?

1.4 Research Objectives

Below mentioned are the research objectives as to what is expected from this study:

- Lyrics sentiment analysis using Natural Language Processing.
- Audio Feature analysis using Machine Learning and Deep Learning Algorithm.
- Building a Recommender System.

All of this exercise is done in order to achieve personalization which in the end enhance the user experience and interaction with the platform which will also be the business aspect of the thesis. Hence this has been tried to achieve by using a combination of machine learning and deep learning models and two separate datasets for lyrics and audio respectively as the combined dataset for both was not available due to piracy concerns. Therefore this is also the major constraint of this thesis as well.

1.5 Report Structure

The report is bifurcated into several sections for a clearer and better understanding of the reader. To begin with the ”Background” understanding of the research topic as to why this research is important in the particular sector or any other sector it affects, ”Research Question” around which the entire report revolves and the ”Objectives” of the report. This is followed by the second section of Literature Review through which the reader will gain enlightenment about the previous work done by other researchers in the same field and this section will also talk about the scope and the significance of this research and will support the research. Then Methodology which is followed to achieve the work which is mentioned in the following section that leads to the next section of Implementation where the steps to achieve the mentioned methodology are explained in details. Furthermore, the Evaluation of the steps mentioned in the above section are verified in the next section called ’Evaluation’. Lastly the research is wrapped by a Conclusion and a Discussion and the Future Scope of the study.

2 Literature Review

2.1 Audio Feature Analysis

Researchers have had a lot of difficulty determining the tempo (or number of beats per minute) of a song or other audio recording in recent years, and this area of study is a popular one right now. Audio synchronization for simultaneous playing has been the topic of numerous applications in the past. (Laroche 2001) discusses methods for identifying the tempo and swing, as well as the beats, in audio recordings, assuming that the tempo is constant. The applications of the paper as discussed in it are synchronising the multiple track audios, automatic looping and so on. The paper has only used beats and tempo with the help of Fast Fourier Transform and has found out the spectral energies. Transient Analysis technique is used to find the upbeat. It is good technique to find the first peak but then the major audio features are missing and also no recommendations are made. Also the algorithm can be very expensive if the length of the song is too long as it only works on the song length.
The automatic classification of audio signals into a hierarchy of musical genres is investigated in this work (Tzanetakis & Cook 2002). Three feature sets are especially suggested for describing pitch content, rhythmic content, and timbral texture. Short Term Fourier Transform is used to calculate MFCC, Zero Crossing amongst others. The entire flow to understand the wavelengths was full wave rectification, low pass filtering then down pass sampling and mean removal to find the autocorrelation. Human survey was conducted to create a bias for understanding but again no recommendations was made for the genres. Although the flow used to find out the autocorrelation is very strong with 61 percent with nonreal time and 44 percent real time success is achieved with 10 music genres. No deep learning methodologies were used for the same. Same dataset as my research study.

Then a research study came (Lee et al. 2009) in which an attempt was made to automatically classify the music genres for which long-term modulation spectral analysis of spectral (OSC and MPEG-7 NASE) as well as cepstral (MFCC) features was put into use. In this study the author majorly focused on the music problem of genre where the labeling of the music tracks is done so that the online categorization of songs is done neatly and efficiently. The study has categorized the audio signals into 3 classes 'short-term' which majorly comprises of timbre features which include mfcc, spectral centroid and others. Long-term is an aggregation of short-term features and Semantic features which refers to human perception like emotions etc. SVM, GMM, LDA, KNN were used to classify the genres. No deep learning models were implied and no recommendations were made.
Post that a study came out to make an MRS in consideration of human emotions (Yoon et al. 2012). For this study a rated list of music preferences of the contestants were taken into consideration. Low-level features of music were used for this study which are known for infusing human emotions. By using automatically extracted low-level elements of music that evoke emotions, this recommendation system was solving the scalability issue of tag-based music recommendation systems. A user history was created by making the contestants fill a form with a couple of personal questions. By listening to that history an attempt was made to reduce the gap between low-level and high level semantic features and this was used to effectively reflect the dramatic changes in the user's behaviour in listening to the song.
In this study (Cheng & Tang 2016), the author developed a hybrid approach for music recommendations that took user personalities and song audio characteristics into account. The author suggested using the support vector machine to address the problems of cold start and sparsity of preference ratings (SVM). User's personal behaviourial information was also used in this study. In this study the personality feature plays an important role in predicting whether the user will like or dislike the song which is compared with the Colaborative Filtering. The problem of cold start is addressed here but is only influenced by the user personal data.
The next study (Chang et al. 2018) also did some interesting research in building a Personal Music Recommendation System (PMRS). The author came up with a recommendation system by using the Convolutional Neural Network with a combination of Colaborative Filtering to pull data from the already existing users. ReLU function was used in CNN but again a pre-existing user history was needed to have a better understanding. Even after using Convolutional Neural Networks no meta data was explored and no proper feature extraction was done.

2.2 Lyrics Dataset

In the first study, the author (Oramas et al. 2018) have generated three separate datasets to evaluate our knowledge extraction strategies. The musical text data was gathered from different websites and then their similarity between their texts was measured by using Ratcliff-Obershelp algorithm. After mapping the texts research map the artists' discography. Words frequencies are also measured. For the information extraction part Natural Language Processing technique is used.

The next study is based on Target-dependent Sentiment Analysis (Gao et al. 2019) but in this study it is not done the traditional way. The author in this study has used Bidirectional Encoder Representations from Transformers (BERT). A sentence level text classification has been performed in this study by using the Target-Dependent BERT. Also the study claims that a well-designed feature engineering with ABSA and BERT can outperform Deep Neural Networks. But again the sentiment analysis done is on plain texts whereas sentiment analysis on songs is very difficult because the interpretation is very subjective person to person. This study (Laurier et al. 2008) is an improvement on the previous NLP studies with text and mood detection because in this the regular music sentiment analysis of a song is mixed with the audio features analysis. Here in this study the lyric are turned into a bag of words and then Lucene Technique is used to find out the similarity between the words. For combining the results for both the audio and lyrics classification there were two approaches, first in which they used separate predictions and then integrating it and the second one was to integrate both the audio and the lyrics features into one vector and place them all in the same location. This made it possible to use audio and song lyrics in the same classifier. Another study comprises of the same approach towards building an intelligent Music Retrieval System. This study also combines audio and text analysis together. The audio features are first extracted like mfcc, spectral centroid using the STFT technique of Fourier Transform. SVM and LDA are used for the analysis. For the lyrics part lexical richness, part-of-speech, bag of words etc. were analysed. This paper implied the same idea but with a very different approach and no recommendations were made. Also no deep learning model was implied.

The literatures above were reviewed on the basis of studies done on attempts to build different forms of MRS. The studies also include some latest works but no study was found which worked on both a combined dataset of lyrics and audio where the song can be downloaded and used for analysis. This is a major concern and a drawback in this study as well, which is the core reason why two separate datasets were used in this study. The above discussed approaches for both the datasets were all ahead of their time and all did a wonderful job and their contributions are invaluable. The gap which should be addressed and no research discussed above was using Natural Language Processing for the sentiment analysis and extracting the audio feature from the scratch. For this study these two tasks are performed with highest accuracy and a recommender system is built using Natural Language Processing with the help of Cosine Similarity and CNN for genre prediction as it is proved superior over other classification models. Also in this study an attempt is made to address the problem of cold start by focusing on the genres of the song and then a new user who plays a song in any one of the genre then the recommendations will be made according to genres closer to one played by the user.

3 Research Methodology

- Dataset 1
  - Performing EDA to achieve a normalised dataset
  - Performing Natural Language Processing Technique by using the sentiment analyser in order to find out the sentiments of the songs.
  - Performing the cosine similarity using the TF-IDF Vectorizer and acquiring a similarity score for the songs which are semantically close to each other.
- Dataset 2
  - Understanding the audio and classifying the genres.
  - 2-D Representation of the sound waves using libROSA Library.
  - Performing the Fourier Transform.
  - Building the Spectrogram and the Mel Spectrogram for the sound waves.
  - Audio Feature Extraction by extracting the timbre features using libROSA Library.
  - Dimensionality Reduction and Normalization.
  - Applying the machine learning and Deep Learning methodology (here CNN is used).
  - Prediction and classification of the genres of the audios selected.

3.0.1 About the Dataset

In this thesis two different type of datasets will be used. The first dataset is the one which only comprises of the lyrics. The dataset has lyrics of 21 different artists.

On the other hand, the second dataset will be used to do the second part of the thesis which is the Audio Feature Analysis. The dataset consists of audio of 30 secs each for different kinds of genres like ‘Rock’, ‘Blues’ and so on.

3.1 Business Understanding

As mentioned in the start of the report that Personalisation is a concept which is being accepted by all types of business modules at a very fast pace may it be online banking and transactions, online shopping and so on. Entertainment is a kind of thing where everyone has their personal taste and preferences. Therefore the industry is focusing on personalisation for a better and smoother user experience. This trend is seen in OTT platforms and all different kind of music platforms as well. Through this study the main focus is that to give the user a personalised music recommendation which he/she likes to listen to. The constraint of the dataset with both the lyrics and the audio is standing in between but research have managed to do them separately. Through the lyric dataset research finding the sentiment giving recommendation and through the audio dataset the audio features are extracted and the recommendation on the basis of the genre of the song is given. Also the problem of cold start which many music platforms face in general is also addressed at an initial level. research have found a way around things as a user history is not required and recommendations on the basis of the genre of the song will be provided. The genre which will be closer to the audio features of the kind of genre the user listens at first will be recommended. Similar thing with the lyric dataset is also done as research have a similarity score for the the songs of 21 artists so recommendations based on that will be done. This entire exercise is revolving just around personalisation which at the end will provide a smoother user experience and will help the users to become loyal users of the so called platforms.

3.2 Data Understanding

For this study there are two different kinds of datasets are used one for the lyrics in which there are different folders of 21 different artists with several songs of them lodged in a csv file. On the other hand the other dataset has wav files of audio for 10 different genres of music. Also the audio dataset contains 2 derived csv file one with the extracted audio features of the music files and the other an augmented file in which those 30 seconds audio files are broken into 3 seconds each and are used for the CNN Model to train the model better.

3.3 Data Cleaning

Because not all models function well with null values, a clean dataset will enhance prediction more accurately and make model prediction easier.

3.4 Data Modelling

The implementation will start after the model has been developed and the dataset has undergone pre-processing. There are 7 Machine Learning Model which are used and apart from that research also have deep learning methodology implied which is Convolutional Neural Network. Which is sometimes known as CNN or ConvNet, is a particular type of deep neural network. A deep, forward-feeding artificial neural network is used. Keep in mind that the fundamental deep learning models, multi-layer perceptrons (MLPs), are also known as feed-forward neural networks.

3.5 Evaluation

How successfully the models accomplish the goals will be assessed in this step. The performance of the model will be assessed using the evaluation matrices below:

3.5.1 Accuracy

It reveals the proportion of times, out of all the classifications in the dataset, where a model correctly identified an item. If the dataset is well-balanced and has high-quality data, accuracy is a useful metric to utilize.

3.5.2 Cosine Similarity

Regardless of the size of the documents, Cosine similarity is a metric used in NLP to gauge how similar they are. The cosine of the angle formed by two vectors projected in a multi-dimensional space is calculated mathematically.

4 Design Specification

Both the datasets that are put to use for this research study is taken from Kaggle. One dataset is purely based on lyrics which will be used to perform NLP and sentiment analysis whereas the other dataset has audio files of 30 secs each from 10 different genres which will be used for genre classification and prediction.

5 Implementation

The research uses different components in order to find out the best recommendation for a new user using the lyrics and the audio features of a song and combine them. The components of the research interact with each other in a sequentially designed manner to get the optimal results. The summary flow of the methodology for both the data sets is given below:

5.1 Data Selection

There are two datasets being used in for this research study one which comprises of the lyrics of the songs from 21 different artists and is used for the sentiment analysis of the lyrics. The second dataset comprises of wav. audio files of 30 seconds each of 10 different genres. Additional to that there are 2 csv files which contains all the derived audio features from the songs. One csv file is the one in which research have 30 seconds audio lodged with their audio feature and also data augmentation is done by breaking the 30 seconds audio to 3 seconds each for training the Neural Network and the machine learning model.
Now the research is divided into two parts in terms of implementation for both the dataset and the flow for both the datasets are different because the objective for both of them are different. [1]

5.1.1 Lyric Dataset

1. Exploratory Data Analysis: The EDA for this dataset consists of majorly tokenization, Lemmatization, finding Unique Word, Data Cleaning, statistics for the visualisation. To elaborate the steps and the flow of the model followed in EDA is that first the lyrics dataset that is being used has so many words which are not even readable to humans let alone the machine to have the model. Also those lyrics have to be converted to words which will be then used as tokens. The unnecessary stopwords are removed from the 'lyrics' as well as the title column. Lemmatization is done to achieve the base root mode of the words and the words are normalized further in order to be used by the machine. This is the essence of the sentiment analysis of this thesis because rather than just finding the semantics of the words in the song the context of the words is at focus in this study to get a better recommendation. After the normalizing of lyrics a simple function to find the unique words is performed to get an idea of the actual words and the number of unique words in the lyrics. Initially the unique list is empty and

[1] Dataset Link: https://www.kaggle.com/datasets/deepshah16/song-lyrics-dataset

the counter will traverse through the lyric's column and then a count of the unique words is done. Those unique words will be added to the newly created 'Word' column. Post this a 'Cleaning' function is defined as to clean the words like 'remix', 'Remix', 'live', 'Live', and so on from the 'Title' column and are stored in a list. Also, if those songs with empty title column is dropped from the list and are not considered. Then a general statistics of the songs for each artist has been done in order to get a clear idea as to what was the count of the words in the entire song and then what is the count coming after the removal of the repeated words.

2. Data Preparation for Visualisation: For this step Lexical Richness is calculated. Lexical richness usually refers to the variety of the words which are being used in the song is calculated so that research can reach to the closest semantic of the word as to in which context the word in the lyrics was used. Before finding the lexical richness, a small step is done in which the length of the song before finding the unique words and after them is done and also the total length of the song is calculated. Then the lexical richness of each song for every artist is calculated. Post this some visualisation for these steps are done. Bar Graph for Unique and Total Word Count, Graph for Lexical Richness, Graph for grouping according to the year and Bar Plot for the individual year is done.

3. Word Cloud: In terms of a song here research is trying to check the number of times a word is appearing in a song for all the songs individually for each artist. Through this the author get an idea that what shall be the target words for each song for each artist.

4. Sentiment Analysis: This is the major part for the lyric dataset. NLP is used to find the sentiment of the songs. Sentiment Analyser is used to find out the sentiments which are categorised into 3 major sentiments namely 'Negative', Positive' and 'Neutral'. A swarm plot is created in which a depiction of the dominant score is predicted. This is very important to do in order to understand the nature of type of lyrics the artists majorly using. Accordingly, the kind of mood set by their music can also be analysed. This is the main objective of this part of the study.

5. Evaluation Metrics: The evaluation matrix used to solidify the findings is Cosine Similarity using TF-IDF Vectorization. A similarity score is given to all the songs and recommendations according to that are made.

5.1.2 Audio Dataset

- Exploratorty Data Analysis: In this part for the audio dataset firstly the 'sample rate' of the audios is calculated for the length of the audios and also frequency through pressure strengths. libROSA library is used to perform the task. The songs usually don't start with a sound and also doesn't end with a sound so a trim leading and trailing silence function is used to remove the abnormal sounds. This will also give us an idea of the starting and ending part of the audio. Post that an initial pictorial 2-D representation is done using libROSA. A song of Reggae genre is used to do the same. This visualization is a time-domain representation of the signal. The author representing loudness (amplitude) against the changing time.

- Data Extraction: The first step in this is 'Fourier Transformation' for better understanding of the visual representation of signals with time constraint and converts them in frequency domain. Here Fast Fourier (FFT) and Short Term (STFT) Transform is used. The former is used because of the discrete nature of the signals and the latter is used to create window lengths of the signals for further analysis. The next step involves creating the Spectrogram and Mel Spectrogram of the signals. Since the author have lost the track of time domain due to FFT, therefore, keeping a track as to which signal is observed first becomes a bit difficult. It will add the logarithmic domain for a better observation of the signals by making us plot the frequency against the time domain. Also the window lengths which were created by STFT, those numbers will be used to keep the track of time as to when the audio starts and where is the peak of frequency on both sides of the axis. Whereas Mel Spectrogram is used to add the Mel scale to normalise the data more and remove the remaining anaomalies for a better analysis of the images.

- Feature Extraction: There are many features which are associated with an audio. The features are extracted. The start is done by finding the zero crossing rate in which a quick padding is done for not having dramatic differences in the sine wave. Next comes Harmonics and Perception which are a kind of multipliers to make those part of the audios audible which are not within the audible range of a human ear. Temo Beat Tracker is used to keep a track of beats with respect to sample rate. Then the main audio features like Spectral Centroid used to find the centre of the spectrum, Spectral Roll Off to find the lowest point of the spectral energy, Mel-Frequency Cepstral Coefficients (mfcc) which is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency and chroma frequencies which project the whole spectrum into 12 bins to represent the 12 unique semitones (or chroma) of the musical octave are done.

These features constitute to our audio features for this research study and the prediction will be done on the basis of their analysis. Some visualization for internal referencing was done by plotting a heat map to find out the correlation between the features and box plot was plotted to find the outliers.

- Dimensionality Reduction and Normalization: Principal Component Analysis (PCA) was done for dimensionality reduction post which MinMax Scaler is used for the noralization process.

- Application of Machine Learning and Deep Learning Models: In this study the author implied 7 machine learning methodologies like GaussianNB, Random Forest, Logistic Regression to name a few and also implemented Convolutional Neural Network (CNN). For this the CSV file in which the audios were broken down to a length of 3 seconds each were used. The model was split into 70 percent train data as more training was required for accurate predictions and the model was tested on 30 percent. After the accuracies calculated for all the machine learning methods and deep learning as well, CNN has the highest rate of prediction of the genre of the audio that was fed to it. To implement CNN, tensor flow was used and 1500 epochs were run and we used Adam Optimizer for the same.

[2]

6 Evaluation

6.1 Evaluating the Methods for lyrics dataset:

For the initial understanding of the data techniques like tokenization, lemmatization, finding unique words and so on was done. A graph for the difference between the total word count and unique word count was plotted as shown below.
This graph was for the clarity of the author as to find out the new words. Post that an analysis was done for the Lexical Richness for the artists as shown in the table below.

Natural Language Processing was used for finding out the sentiment score for all the songs of each artists. A total of three categories were formed namely ’Positive’, Negative’ and ’Neutral’. A dominant score after analysing each song on these three emotions was derived. Through the above table it is clear that every

[2] Dataset Link: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

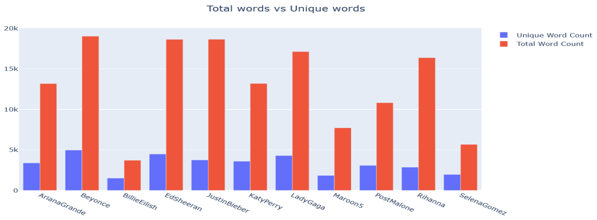
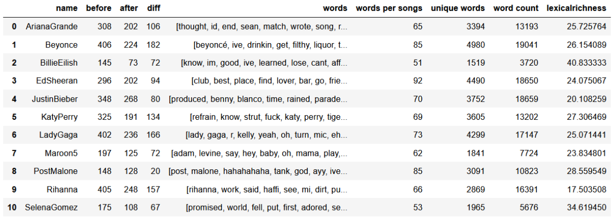
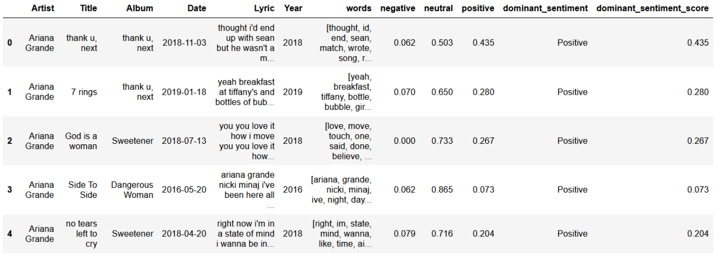


Figure 1:
Word Count of Unique vs Total Words



	name	before	after	diff	words	words per songs	unique words	word count	lexicalrichness
0	ArianaGrande	308	202	106	[thought, id, end, sean, match, wrote, song, r...	65	3394	13193	25.725764
1	Beyonce	406	224	182	[beyoncé, ive, drinkin, get, filthy, liquor, t...	85	4980	19041	26.154089
2	BillieEilish	145	73	72	[know, im, good, ive, learned, lose, cant, aff...	51	1519	3720	40.833333
3	EdSheeran	296	202	94	[club, best, place, find, lover, bar, go, frie...	92	4490	18650	24.075067
4	JustinBieber	348	268	80	[produced, benny, blanco, time, rained, parade...	70	3752	18659	20.108259
5	KatyPerry	325	191	134	[refrain, know, strut, fuck, katy, perry, tige...	69	3605	13202	27.306469
6	LadyGaga	402	236	166	[lady, gaga, r, kelly, yeah, oh, turn, mic, eh...	73	4299	17147	25.071441
7	Maroon5	197	125	72	[adam, levine, say, hey, baby, oh, mama, play,...	62	1841	7724	23.834801
8	PostMalone	148	128	20	[post, malone, hahahahaha, tank, god, ayy, ive...	85	3091	10823	28.559549
9	Rihanna	405	248	157	[rihanna, work, said, haffi, see, mi, dirt, pu...	66	2869	16391	17.503508
10	SelenaGomez	175	108	67	[promised, world, fell, put, first, adored, se...	53	1965	5676	34.619450

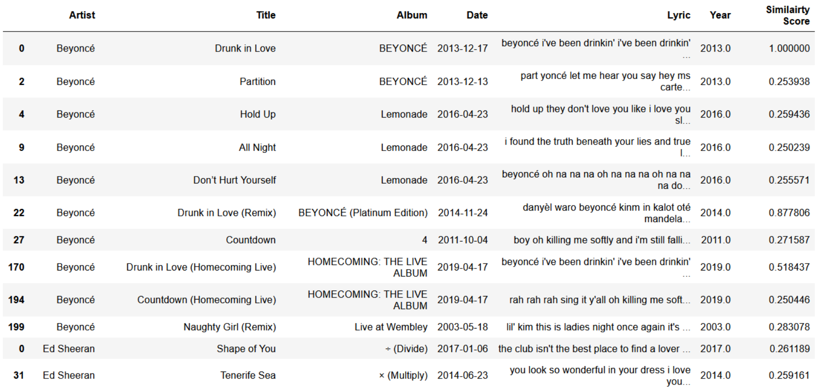
Figure 2:
Lexical Richness Table



	Artist	Title	Album	Date	Lyric	Year	words	negative	neutral	positive	dominant_sentiment	dominant_sentiment_score
0	Ariana Grande	thank u, next	thank u, next	2018-11-03	thought i'd end up with sean but he wasn't a m...	2018	[thought, id, end, sean, match, wrote, song, r...	0.062	0.503	0.435	Positive	0.435
1	Ariana Grande	7 rings	thank u, next	2019-01-18	yeah breakfast at tiffany's and bottles of bub...	2019	[yeah, breakfast, tiffany, bottle, bubble, gir...	0.070	0.650	0.280	Positive	0.280
2	Ariana Grande	God is a woman	Sweetener	2018-07-13	you you love it how i move you you love it how...	2018	[love, move, touch, one, said, done, believe, ...	0.000	0.733	0.267	Positive	0.267
3	Ariana Grande	Side To Side	Dangerous Woman	2016-05-20	ariana grande nicki minaj i've been here all ...	2016	[ariana, grande, nicki, minaj, ive, night, day...	0.062	0.865	0.073	Positive	0.073
4	Ariana Grande	no tears left to cry	Sweetener	2018-04-20	right now i'm in a state of mind i wanna be in...	2018	[right, im, state, mind, wanna, like, time, ai...	0.079	0.716	0.204	Positive	0.204

Figure 3:
Sentiment Analysis by dominant score

song gained a dominant sentiment and score according to which it was analysed and recommendations were made. Cosine Similarity was used for the evaluation of the NLP model and the first song in the lyric dataset 'beyoncé i've been drinkin' i've been drinkin' by the artist Beyonce was used to find the similar songs to it. A range of 0.25 was used for the cosine similarity to predict the similarity and recommendations are made. The list of the songs which are shown in the table are a result of the similarity score based on the sentiment of the song used for checking the similarity. The recommendations made on were of good accuracy and was showing a decreasing trend in the table. A user who is new and comes and plays a song online can be suggested songs on the basis of lyrics in this fashion.



	Artist	Title	Album	Date	Lyric	Year	Similairty Score
0	Beyoncé	Drunk in Love	BEYONCÉ	2013-12-17	beyoncé i've been drinkin' i've been drinkin' ...	2013.0	1.000000
2	Beyoncé	Partition	BEYONCÉ	2013-12-13	part yoncé let me hear you say hey ms carte...	2013.0	0.253938
4	Beyoncé	Hold Up	Lemonade	2016-04-23	hold up they don't love you like i love you sl...	2016.0	0.259436
9	Beyoncé	All Night	Lemonade	2016-04-23	i found the truth beneath your lies and true l...	2016.0	0.250239
13	Beyoncé	Don't Hurt Yourself	Lemonade	2016-04-23	beyoncé oh na na na oh na na na oh na na na do...	2016.0	0.255571
22	Beyoncé	Drunk in Love (Remix)	BEYONCÉ (Platinum Edition)	2014-11-24	danyèl waro beyoncé kinm in kalot oté mandela...	2014.0	0.877806
27	Beyoncé	Countdown	4	2011-10-04	boy oh killing me softly and i'm still falli...	2011.0	0.271587
170	Beyoncé	Drunk in Love (Homecoming Live)	HOMECOMING: THE LIVE ALBUM	2019-04-17	beyoncé i've been drinkin' i've been drinkin' ...	2019.0	0.518437
194	Beyoncé	Countdown (Homecoming Live)	HOMECOMING: THE LIVE ALBUM	2019-04-17	rah rah rah sing it y'all oh killing me soft...	2019.0	0.250446
199	Beyoncé	Naughty Girl (Remix)	Live at Wembley	2003-05-18	lil' kim this is ladies night once again it's ...	2003.0	0.283078
0	Ed Sheeran	Shape of You	÷ (Divide)	2017-01-06	the club isn't the best place to find a lover ...	2017.0	0.261189
31	Ed Sheeran	Tenerife Sea	× (Multiply)	2014-06-23	you look so wonderful in your dress i love you...	2014.0	0.259161

Figure 4:
Similarity score with value of similarity score¿0.25

6.2 Evaluation of Methods used for Audio Dataset

The audio dataset was comparitively larger and more complex to handle. The general flow of the model was to first represent the signals in 2-D form using Fourier Trasform by using the sample rate. Then FFT and STFT with hop length and Window Size and the signals were changed to frequency domain from time. Post that the signals were converted to spectrogram and Mel spectrogram as the time

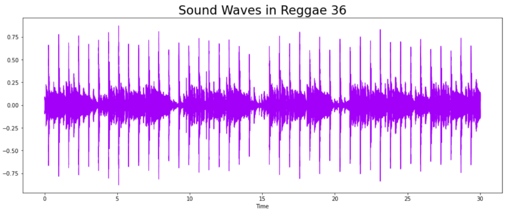


Figure 5:
2-D Representation

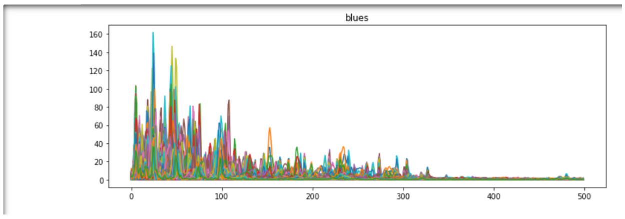


Figure 6:
Fourier Transform

constraint was removed so window length was used to keep track. FFT Window size was kept 2048 and the shape was printed as 1025, 1293. The darker sections were the one with higher frequencies. Post that Mel scale was added and Mel Spectrogram was plotted because of some non-linear transformations. Spectral-Roll, Spectral centroid, mfcc were among some timbre features. The model was fit with train at 70 percent and 7 Machine Learning Methodologies were used with their accuracies shown below.

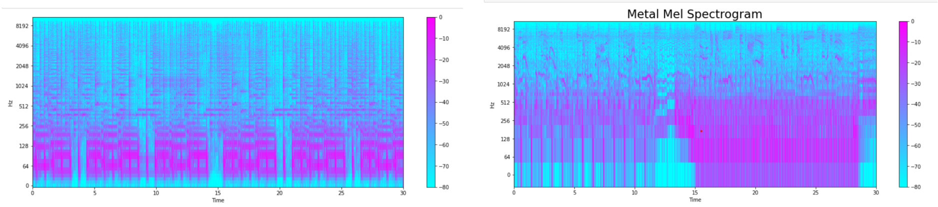


Figure 7:
Spectrogram (Up) and Mel Spectrogram (Down)

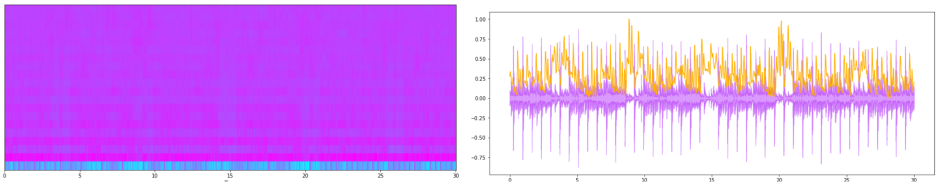
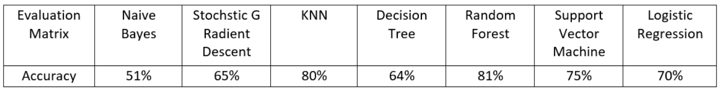


Figure 8:
MFCC (Up) and Spectral Centroid (Down)

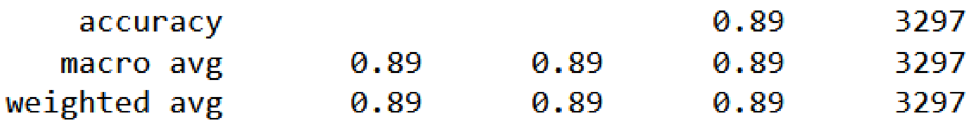


Evaluation Matrix	Naive Bayes	Stochstic G Radient Descent	KNN	Decision Tree	Random Forest	Support Vector Machine	Logistic Regression
Accuracy	51%	65%	80%	64%	81%	75%	70%

Figure 9:
Accuracy of Machine Learning Models Implied

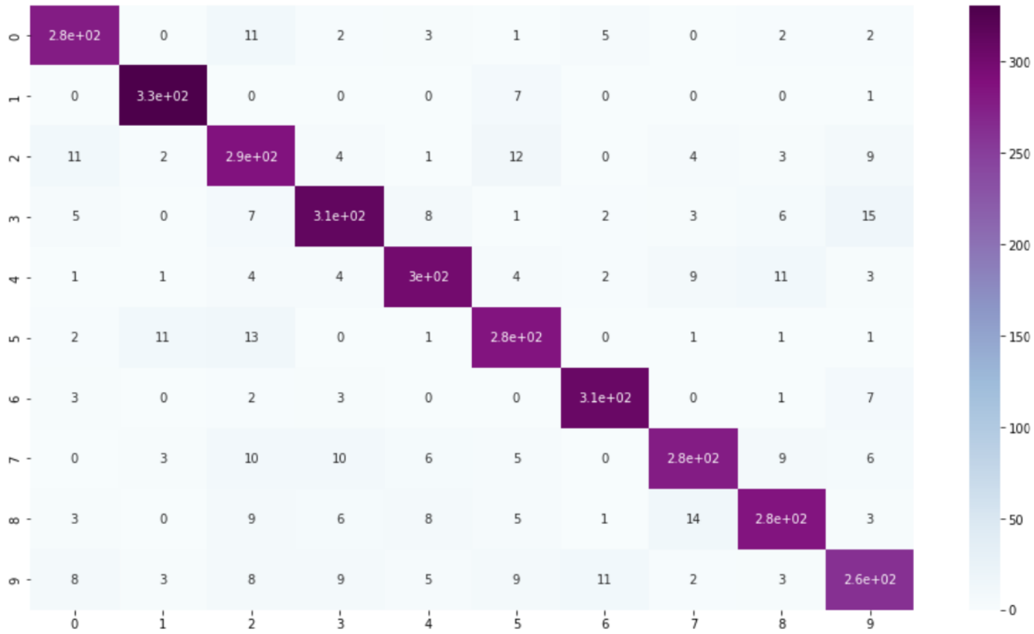
From the above table it can see that Random Forest had the best accuracy but this was not it. Post this the XGB Classifier is also applied on training the dataset for the classification and prediction.
The accuracy of XGB is higher than all the 7 model which comes out to be 89 percent.

A Confusion Matrix was plotted to find the correlation between all the ML models before CNN was applied.



```
     accuracy                           0.89      3297
    macro avg       0.89      0.89      0.89      3297
 weighted avg       0.89      0.89      0.89      3297
```

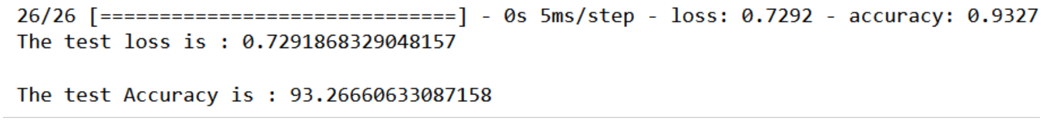
Figure 10:
Accuracy of XGB



	0	1	2	3	4	5	6	7	8	9
0	2.8e+02	0	11	2	3	1	5	0	2	2
1	0	3.3e+02	0	0	0	7	0	0	0	1
2	11	2	2.9e+02	4	1	12	0	4	3	9
3	5	0	7	3.1e+02	8	1	2	3	6	15
4	1	1	4	4	3e+02	4	2	9	11	3
5	2	11	13	0	1	2.8e+02	0	1	1	1
6	3	0	2	3	0	0	3.1e+02	0	1	7
7	0	3	10	10	6	5	0	2.8e+02	9	6
8	3	0	9	6	8	5	1	14	2.8e+02	3
9	8	3	8	9	5	9	11	2	3	2.6e+02

Figure 11:
Confusion Matrix for ML Models

After this Deep Learning methodology was used and CNN was used for classification as CNN over powers other classification Model. This is what exactly happened as CNN used tensor flow and had the highest accuracy of 93 percent. Also CNN was used because audio signals were converted into images and CNN works best with image dataset. Research have used the ReLU activation for the model. The model is predicting and classifying the genres 93 times out of 100 which was convincing rate for an unsupervised learning model. Adam Optimiser was used for this purpose along with Tensor Flow and through this an attempt to solve the problem of cold start was also done.



```
26/26 [==============================] - 0s 5ms/step - loss: 0.7292 - accuracy: 0.9327
The test loss is : 0.7291868329048157

The test Accuracy is : 93.26660633087158
```

Figure 12:
Accuracy for CNN

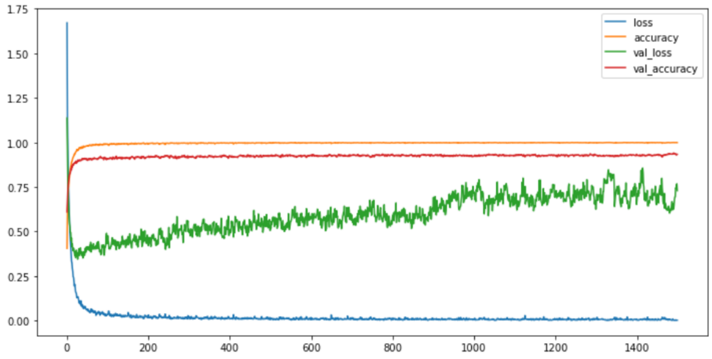


Figure 13:
Accuracy Curve for CNN

7 Conclusions and Discussion

After a thorough research online and offline the main motive of this study was to create a strong recommender system which actually just not focus on the semantics of the lyrics just for the sake of it. Music is something which is conveyed beyond words. The real challenge in building this recommender system was to actually just not focus on the semantics of the words but to also get the context of the words which are used in a song. For this an attempt has been made in this study to catch the context and three different moods were identified for the lyrics. Also

a similarity score using cosine similarity with the help of TF-IDF vectorization was done. Also, when it comes to music the audio also plays an important part, therefore an attempt was made to understand the audio features of the songs and to extract the major features and make the machine understand them in a pictorial way so that research finds out which genres were closer to each other. For the lyric dataset a sentiment score based on the words or text was given to each song and then using a similarity score a recommendation which is closest to the first song is done. For the audio dataset CNN model is able to predict the genres for 93 percent of the time when an audio is given to it (the accuracy for other models is mentioned in the table above). Also on the other hand for the sentiment analysis part in the lyrics dataset the accuracy was very close and the mood prediction using cosine similarity was coming out to be quite convincing.
Now the main problem which was being faced by these kinds of MRS was of cold start which was also a big business problem indeed. The problem of cold start in a layman terms means that there is no data to start with for the user. You have to create a user data first and then the machine can make the recommendations. But then a pre-existing user history is not possible for a new user. Hence in this study the author have found a way around that which is the author have created a pre-existing history of the genre of the songs by extracting the audio features of the same. Now if a person plays a song then research already have a list of genres in a decreasing order of the similarity which can be and will be recommended to the user. Also, these were the main objectives of the research study at large which were mentioned earlier in the start. The attempt is made to fulfill the objectives and come up with an idea which actually promotes the feeling of personalization and enhances the user experience at all levels through an efficient MRS because music does play an important part in everyone's life at some point of time. This is at very basic level but can be considered a good start also in terms of business perspectives as through this the user experience and interaction can be enhanced.

8 Future Scope

Although the accuracy of the CNN Model came out to be 93 percent, the biggest constraint in this study was to find a dataset with both lyrics and audio available together. By that a much better, efficient and personalised Music Recommendation System can be built. The problem of cold start can be then further solved without a pre-existing user history. Many music platform owning companies can use this model and improvise it for a much more smoother user experience.

Acknowledgement

I would want to thank my parents, my teachers, and my friends for providing support and direction in preparing the research.

References

Chang, S. H., Abdul, A., Chen, J. & Liao, H. Y. (2018), A personalized music recommendation system using convolutional neural networks approach, Institute of Electrical and Electronics Engineers Inc., pp. 47–49.

Cheng, R. & Tang, B. (2016), A music recommendation system based on acoustic features and user personalities, Vol. 9794, Springer Verlag, pp. 203–213.

Gao, Z., Feng, A., Song, X. & Wu, X. (2019), 'Target-dependent sentiment classification with bert', *IEEE Access* **7**, 154290–154299.

Laroche, J. (2001), Estimating tempo, swing and beat locations in audio recordings, pp. 135–138.

Laurier, C., Grivolla, J. & Herrera, P. (2008), Multimodal music mood classification using audio and lyrics, pp. 688–693.

Lee, C. H., Shih, J. L., Yu, K. M. & Lin, H. S. (2009), 'Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features', *IEEE Transactions on Multimedia* **11**, 670–682.

Oramas, S., Espinosa-Anke, L., Gómez, F. & Serra, X. (2018), 'Natural language processing for music knowledge discovery', *Journal of New Music Research* **47**, 365–382.

Tzanetakis, G. & Cook, P. (2002), 'Musical genre classification of audio signals', *IEEE Transactions on Speech and Audio Processing* **10**, 293–302.

Yoon, K., Lee, J. & Kim, M. U. (2012), 'Music recommendation system using emotion triggering low-level features', *IEEE Transactions on Consumer Electronics* **58**, 612–618.