Emotion Integrated Music Recommendation System Using Generative Adversarial Networks

Mrinmoy Bhaumik[1], Patricia U Attah[1],, Dr. Faizan Javed,
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
mbhaumik@mail.smu.edu
pattah@mail.smu.edu
abehuriapathak@mail.smu.edu

Abstract Music has the power to stimulate emotions within us hence is often called the "language of emotion," because of this there has been scientific research done on how music affects our emotions. keeping this in mind this study attempts to explore the use of emotion as a key feature in generating a playlist with a deep learning model to improve upon the current music recommendation system. This study will sample emotions from certain subjects for each song in a sample of the data. Due to subjectivity of emotion, must combine the emotions from more than one individual to get the perceived emotion, a semi-supervised model will then be used to predict the emotion for the unsupervised data. A content-based recommendation system will be built using a GAN (Generative Adversarial Network). **'Result Sentences placeholder'**

1. Introduction

Advances in technology has given us an information overload and web-based music services can employ recommender systems as a powerful tool to harness this information and create a better experience for the user. There are different recommendations systems currently in use most of them use features like user history, popularity, danceability and other metrics to generate recommendations, more advanced systems use collaborative or content-based systems. Considering the significant effect music has on emotion we intend to investigate the use of content-based recommendation system by incorporating the emotional aspect of the song with the advantages of using a generative adversarial model.

Recommender Systems (RS) can be classified into three groups Collaborative Filtering (CF), Content-Based (CB) and a Hybrid method which incorporates both prior named methods. CF relies on how other users rated the item and compares the similarities between different users and recommends items to other based on the similarity of their rating. There are two approaches to CF, which is model-based, and memory-based CF. Memory-based CF uses the user rating historical data to compute calculate similarity between users or items. There are two varieties of memory-based CF which includes: item-based and user-based, where item-based looks for users who liked comparable items and output recommendations based on items those users liked, while user-based finds similar users based on their ratings of items and recommends items those users liked. Memory based approach uses underlying algorithms such as cosine similarity or pearson correlation. Model-based CF uses machine learning algorithms to predict users' ratings of unrated items. Some methods include matrix factorization, clustering, and deep learning. Content-based systems do not require user ratings at all and instead makes recommendations using the features of the items from the user's history. Considering the significant effect music has on emotion the purpose of this study is to investigate the use of content-based recommendation system by incorporating the emotional aspect of the song.

Current music recommender systems have not leveraged the class of emotion conveyed in music such as Apple music, Pandora, Amazon music and Spotify. Spotify employs three types of recommendation systems collaborative filtering, natural language processing and audio file models [15]. Their collaborative models recommend songs by collecting information about what songs other users like, the recommendations are based on series of features determined for each song, as well as user history. In previous studies there have been recommender systems built with a variety of deep learning models but there have not been any that utilized emotion as a key feature in combination with the advantages of an adversarial generative model. There has been a study that used a Convolutional Neural Network (CNN) for emotion-based recommendations, however this differs from this study using a GAN as the framework of GANs has an built-in self-checking mechanism which attempts to detect whether an item is from the sample data i.e., the users original track list or not. This unique trait of GAN will better help in identifying which music a specific user would prefer.

A GAN is a generative model using deep learning techniques such as neural networks. Gans are frameworks consisting of two sub-models are known as the generator, and the discriminator. The generator creates similar data to the training data while the discriminator attempts to determine which data is real (from the training data) and which is fake (from the generator). Prosvetov [7] did a study where he created GAN recommender system where the system would recommend airline tickets to customers and compared it to a recommender system that was based on a Deep Neural Network and was able to successfully compete with it.

Most recommender systems suffer from data sparsity which occurs because users only interact with a small portion of the objects when being used with matrix operations, this can be overcome with GAN since the premise of the model is not matrix operation but just reproducing synthetic samples, another problem that recommender systems face is data noise this is negative or misleading samples that are uninformative and cause inaccurate results this particular situation rarely applies for this case since the music selection of an individual should not be wrong or inaccurate however if someone does have a song in their playlist they do not like a GAN will help determine this since the main function is to identify the true distribution of the selection of songs.[11]

Results Paragraph

Conclusion Paragraph

2. Related Works

Authors have studied the music recommendation systems in context to emotions with different approaches. [8] In prior research authors have investigated music track selection to change or maintain emotional and psychophysiological states to support mental wellbeing. The study emphasizes on the use of music influence on humans, by attempting to identify the current mental state the individual is in by finding emotional and other features in the music in combination with external factors, personal data, and behaviors to alter their mental state. This study collects a wide variety of data such as the current state they are in, general personal data, physical sensory data, external factors such as social network-based applications, interactive user feedback as the individual listens to the track. With all this data collection it makes it difficult in narrowing down what if any of these factors influenced the individual's emotional state. This angle chooses to alter emotional state and is not aiming to generate a satisfying experience but a chosen emotional state end point. There have also been recent studies have investigated facial expression to determine emotional states which can be used to generate a playlist. The results were not expressed in mathematical terms of error or otherwise but were just a comparison of similar selected features as was assigned by their feature detection software. But they reported the system was successful in determining the mood changes during listening to music. [9] This study assumes that if the facial expression can be properly identified into one of the following categories happy, sad, fear, anger, surprise, or disgust this would generate an engaging playlist, which is another interesting approach. [12] There has been research on emotion-based music recommendations for films, in this study they investigated soundtracks of films and generated emotions from the music features and film video where captions, sound effects, visual features and speech can also be factored in. This study seeks to create certain emotions that can be used for articulate film expression. In this study they use a modified Mixed Media graph (MMG) to extract the emotion features and a Music Affinity Graph (MAG) to discover the relationship between the music features and the emotions. The result was that the best algorithm used between the two models they compared was 0.85 percent accurate.

Researchers in [1] have also used a music perception model to detect the emotion of audio file in terms of valence and arousal index, this was done by continuously observing songs aired on popular radio stations and creating radio induced emotion dataset. Back then, Radio music was an effective way of inducing emotions to influence decision making for marketing/Selling products.

Music emotion recognition was done by feature selection method, by selecting features that contribute most to the representation of music, acoustic features were the most prevalent in the feature selecting procedure. Arousal–valence emotional plane is a continuous space. Each point of the emotion plane represents an emotion recognized by the regression model of VA. Regression equation is used to predict an unknown dependent variable, which is established by determining the correlation between the dependent variable and some independent variables. The linear regression theory was used to predict (V, A) values of the music in the emotional plane.

There are also examples of studies where emotion is studied in other mediums related to music. There has been research on emotion-based music recommendations for films, in this study they investigated soundtracks of films and generated emotions from the music features and film video where captions, sound effects, visual features and speech can also be factored in. This study seeks to create certain emotions that can be used for articulate film expression [10].

Ashu Abdul, Jenhui Chen, Hua-Yuan Liao and Shun-Hao Chang [2] proposed an emotion-aware personalized music recommendation system (EPMRS) to extract the correct song based on the mood. This system is a combination of the deep convolutional neural networks (DCNN) approach and the weighted feature extraction (WFE) approach. The DCNN approach helped to extract the latent features from music data for classification. while in the WFE approach, implicit user ratings are generated for the music are generated to extract the correlation between the user data and the music data. The system recommends the songs to the user based on calculated implicit user rating for the music.

In related research authors used a large dataset and weighted feature extraction from user-to-song relationships that are determined from user data. A deep convolution neural network (CNN) is deployed to get the song's latent music features. The user's listening history and audio signals from the specific songs to help classify and recommend songs to the user. Overall, the study suggests that there were some positive outcomes from this type of recommender system, but it would also need to continuously assess what the user is listening to it at the moment.This research heavily relies on the CNN to do this type of classification and it would be interesting to see if this classification can be done more effectively with another type of deep learning model.

In [20], the objective of the research is a way to classify emotions through music using SVM. Emotional states considered anger, sadness, and happiness. Subjectivity when it comes to emotion classification with music is a problem, to reduce the issue they build a classification system applied to different subjects. This is also a project which uses subjects to determine emotion. Subjects were asked to assign an emotion after hearing a certain song and then there were total of 24 emotion cognitive tracks made for each subject. There are two SVMs that are created, one for arousal and one for valence. This can be an effective way to classify music within the main emotional states. One of the ways to determine the strength of their model was to use another music recommendation system approach to see the validity of their method. It seemed to compete well with the other systems, but the only issue was that there needed to be a larger collection music to pull from to get a more accurate model.

3. Data

The data used was originally sourced from Spotify, downloaded from Kaggle. It consists of 25 features and 18,454 instances of songs. Significant audio features were extracted from the data. When identifying emotions in music it is prone to subjectivity as stated in past articles different people perceive emotions differently. The main objective of this paper is to find out how effective the emotion conveyed in a song will help in generating a successful playlist, therefore this study does not delve into the uncertainty of generating emotion from audio features, environment, or user interface activities because these sources may not convey true emotion that is being perceived by users and needs to be authenticated. In our study the emotions from various subjects are collected and aggregate this emotion down to either one or two primary emotions conveyed by the song. The emotion sampling capability is limited, and it is not possible to obtain emotions for every song, so a semi-supervised machine learning model is used to complete the process.

3.1 Data Exploration

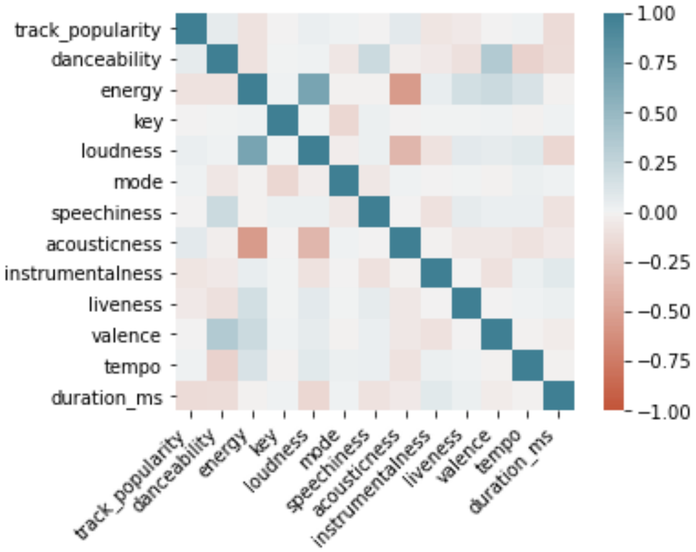

Fig. 1 The above figure is a correlation plot between the numerical features in the Spotify dataset.

Going through the Spotify dataset that was made available through the Spotify API there is a special focus on valence. Valence is the measure of positivity in a track where a high valence means the song is happy, cheerful, or euphoric while a low score would be considered sad, depressed, or angry. From Fig 1. The plot shows there is not much correlation with valence and the other numerical variables. The only two other features that show a correlation around a .5 with valence are energy and danceability. So, looking at the energy and danceability specifically compared to valence one can see a positive linear regression with both.

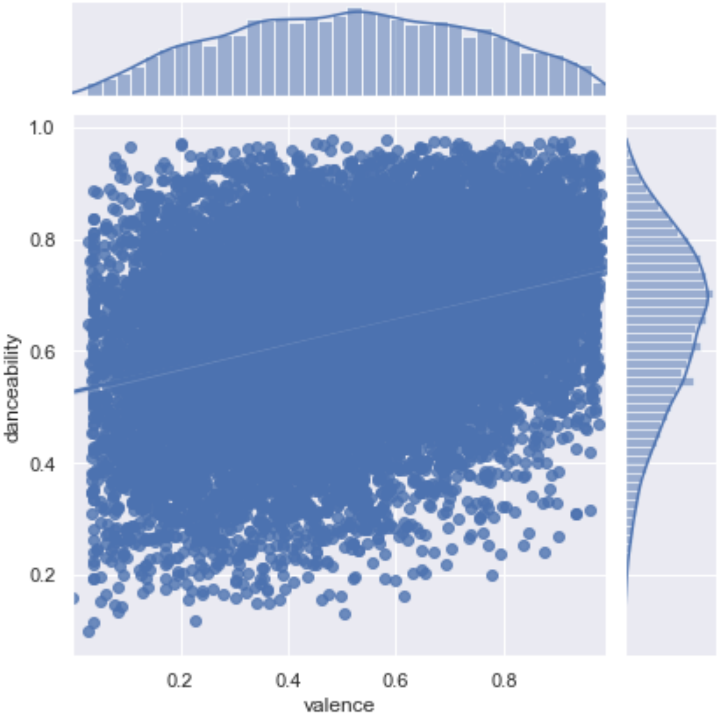

Fig. 2 The above figure is scatterplot of the danceability and valence features of the Spotify data showing that there is a positive correlation.

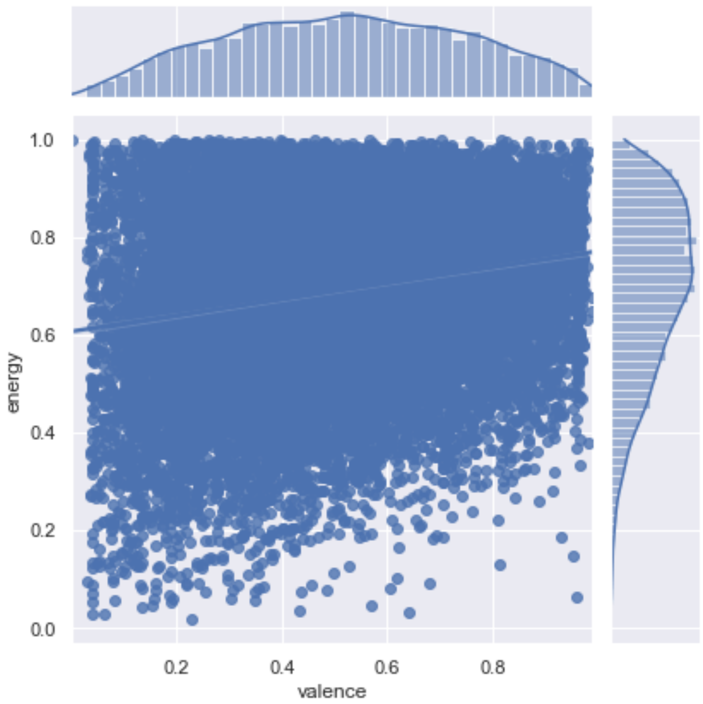

Fig. 3 The above figure is scatterplot of the energy and valence features of the Spotify data showing that there is a positive correlation.

There were 6 major playlist genres which are rock, r&b, pop, edm, latin and rap, these were divided into playlist sub-genres which are 24 in number. Genre is a significant feature as from initial analysis there seems to be a clear distinction between the playlist genre and various features in the data set including valence, danceability, energy, acousticness and loudness. There is a large variation in the median acoustic values between genres, r&b has the highest median value of acousticness at about 0.18 followed by Latin music at 0.17 while EDM and rock has the lowest values at about 0.12. The acousticness measures how confident that the song is acoustic. The genres with highly danceable music are rap, latin and r&b with median values of about 0.71, 0.7 and 0.68 respectively, while rock had an exceptionally low danceable value of 0.5. The genres have similar valence levels except latin which has a high median value of 0.62, while edm has a low median value of 0.38.

Investigating individual scatter plots a slightly negative correlation between energy and the how acoustic the song was as the energy increased the acoustic level slightly decreased, and slightly positive correlation between the loudness and the energy. Also, it was noticed that speechiness, liveliness and acousticness were strongly right skewed, while loudness was strongly left skewed depending of the method of analysis used these variables may have to be transformed.

4. Methodology

1. Sample data for emotion from subjects.
2. Aggregate sampled emotions.
3. Use semi-supervised model to predict emotion in unsampled songs.
4. Generate playlist using generative adversarial networks.

Content-based recommendation models

- Item-based nearest neighbors
- GAN (Generative Adversarial Network)

Potential Additional steps would be to perform natural language processing on lyrics of the song to add another sentiment to the data features being analyzed.

4.2 Emotion labeling using Semi Supervised Models

The data used for this project is a combination of two data sets one data set has 686 rows and 19 columns, while the other data consists of 18,454 rows and 25 columns. The smaller data set has emotion identified; however, this data set does not have the genre of the song. While the larger data set that will be concatenated to the first does not have emotion classified. In this study there will be several sample emotions that will be obtained by sending a survey that will have people identify the emotion they experienced by the song this will be aggregated and a final emotion will be determined by the percentage of an emotion expressed. Due to limited time and resources this cannot be completed for the whole data set so only 10% of the songs in dataset will be labeled and a semi supervised model will then be used to predict the rest of the data.

Semi-supervised learning makes use of both labeled and unlabeled training data. The premise behind the labeling is label propagation where an assumption is made of an edge connecting two nodes together carry a notion of similarity, this assumption can be made since items with similar feature values will tend to connect.

4.2 Generative Adversarial Networks

Generative adversarial networks (GAN) are a framework consisting of two neural networks a generative and discriminative network, these two models are trained concurrently. The generative model uses unsupervised learning to automatically discover patterns and trends in the data that can capture the data distribution to generate new instances of the data, while the discriminative model estimates the probability of the data coming from the actual data set or the generated one. The overall idea is that these two models are competing, the generator is creating synthetic music features that it passes to the discriminator with hopes to maximize the probability of the discriminator not identifying that the data is synthetic. The goal of this competition is to improve the model until the model generated music is for an individual is indistinguishable from the music an individual would have listened to, hence creating an improved recommender system.

The generative model applied in this system is a CNN. In prior works the input would be two sparse vectors one for the users the other for the item. Embeddings are created for both user and items. The user embeddings v(u) is created by mapping the user to its look up table, while the item embedding v(i) is created by mapping the item to the item look up table it is associated with then passed to a Long Short Term Memory model (LSTM) to generate a multi-dimensional embedding for each item. In this paper however the audio features have already been extracted, therefore the data can be passed in as is since the audio features will serve as the embedding values. The resulting representations v(u) and v(i) are concatenated by the multilayer perceptron. The training is performed using the users and the music features once the architecture has been trained it would be able to predict the songs a user would like based on rankings of the songs. The song rankings that because of the output from the recurrent neural network will be what is accessed by the discriminator.

In summary the GAN consists of two networks:

1. A generative network G, that takes input pz(Z) with input z with a density pz and returns output xg = G(z)
2. A discriminative network D that takes the input x which may either be ‘true’ (xt with a density of pt) or ‘generated’ (xg with a density of pg, which is the result of density pz going through G) and returns a probability D(x) of ‘true’ against generated.

The GAN system works better when both frameworks are neural networks. A prior is defined on the input noise variables pz to learn the generators distribution pg over the data x. then map the data to the space called Gz where G is a differentiable function. The following equation depicts the absolute error of the discriminator.

$$E(G,D)=\frac{1}{2}E_{x-p_t}[1-D(x)]+E_{x-p_g}[D(z)]$$

References

1. Panwar, S., Roopaei, M.,Rad,P., Choo, K. (2019). Are you emotional or depressed? Learning about your emotional state from your music using machine learning. Journal Of Computing. Vol. 75 Issue 6, p2986-3009. 24p.
2. Ashu A., Chen, J., Hua-Yuan, L., Shun-Hao C. (2018). An Emotion-Aware Personalized Music Recommendation System Using a Convolutional Neural Networks Approach. **Applied Sciences; Basel** Vol. 8, Iss. 7
3. Oramas, S., Nieto, O., Sordo, M.,Serra,X. (2017). A Deep Multimodal Approach for Cold-start Music Recommendation. ACM Proceedings of the 2nd Workshop on deep learning for recommender systems, 2017-08-27, p.32-37
4. Yepes, Fabio A ; López, Vivian F ; Pérez-Marcos, Javier ; Gil, Ana B ; Villarrubia, Gabriel (2018). Listen to This: Music Recommendation Based on One-Class Support Vector Machine. Cham: Springer International Publishing Hybrid Artificial Intelligent Systems, 2018-06-08, p.467-478
5. Vall, Andreu ; Widmer, Gerhard (2019). Machine Learning Approaches to Hybrid Music Recommender Systems. Cham: Springer International Publishing Machine Learning and Knowledge Discovery in Databases, 2019-01-18, p.639-642
6. Antal D.,Fletchter A., Ormosi L. P. (2021) Music Streaming: Is it a Level Playing field? Competition Policy
7. Prosvetov A. V. (2019) GAN for Recommendation System, Journal of Physics: Conference Series
8. Xinxi Wang and Ye Wang. 2014. Improving Content-based and Hybrid Music Recommendation using Deep Learning. In Proceedings of the 22nd ACM international conference on Multimedia (MM '14). Association for Computing Machinery, New York, NY, USA, 627–636. DOI:https://doi.org/10.1145/2647868.2654940
9. H. Zhang, H. Yang, T. Huang and G. Zhan, "DBNCF: Personalized Courses Recommendation System Based on DBN in MOOC Environment," 2017 International Symposium on Educational Technology (ISET), 2017, pp. 106-108, Doi: 10.1109/ISET.2017.33
10. van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), Advances in Neural Information Processing Systems 26 (2013) (Vol. 26). Presented at the Neural Information Processing Systems Conference (NIPS 2013), Lake Tahoe, NV, USA: Neural Information Processing Systems Foundation (NIPS).
11. Min Gaoa, Junwei Zhanga, Junliang Yuc , Jundong Lid, Junhao Wena, and Qingyu Xiong (2020). Recommender Systems Based on Generative Adversarial Networks: A Problem-Driven Perspective, Key Laboratory of Dependable Service Computing in Cyber Physical Society (Chongqing University), Ministry of Education, Chongqing, 401331, China
12. Shan, M.-K., Kuo, F.-F., Chiang, M.-F., & Lee, S.-Y. (2009). Emotion-based music recommendation by affinity discovery from film music. *Expert Systems with Applications*, *36*(4), 7666–7674. https://doi.org/10.1016/j.eswa.2008.09.042
13. H. Immanuel James[1], J. James Anto Arnold[2], J. Maria Masilla Ruban[3], M. Tamilarasan[4], R. Saranya[5] (2013) Emotion based music recommendation system e-ISSN: 2395-0056, p-ISSN: 2395-0072
14. Mikhail Rumiantcev, Oleksiy Khriyenko, Emotion Based Music Recommendation System, ISSN 2305-7254