APS360 PROJECT PROPOSAL

Jonas Martins
Student# 1006869907
jonas.martins@mail.utoronto.ca

Jeongwoong (Daniel) Choi
Student# 1004942743
jeongwoong.choi@mail.utoronto.ca

Jennifer Sunny
Student# 1006998732
jennifer.sunny@mail.utoronto.ca

Chielotam Agbatekwe
Student# 1006988057
chielotam.agbatekwe4@mail.utoronto.ca

ABSTRACT

This project analyzes the profound influence of music on human emotions and behaviors, aiming to utilize deep learning for music emotion analysis. By categorizing music into distinct moods, the initiative seeks to enable users to curate playlists that resonate with their emotional states, enhancing their listening experience. The research involves both traditional Machine Learning and Deep Learning methodologies to recognize and classify the emotions evoked by music. In our report, we discuss the intricacies of Music Emotion Recognition (MER), highlighting the significance of various musical aspects such as rhythm, harmony, tempo and pitch. We lay out a cohesive data processing method that extracts four feature spectrograms from raw music data, and model architecture consisting of a Convolutional Neural Network (CNN) followed by a Recurrent Neural Network (RNN). The project also emphasizes the ethical considerations, particulary concerning copyright issues, and outlines a comprehensive project plan to ensure efficient collaboration and timely completion. Our goal is to harness the power of deep learning to provide a refined understanding of music's emotional impact, bridging the gap between music, technology, and human emotion.

—-Total Pages: 8

1 INTRODUCTION

Music has a significant influence on human emotions and behaviors. Exposure to certain melodies, rhythms, harmonies amongst other things can help alleviate bad moods and increase productivity. Music can also serve as a means of expression and a way to find community and understanding. The goal of this project is to employ music sentiment analysis to categorize popular music into genres or moods. The practical application of this is to grant a user the ability to curate playlists based on their desired emotional state in hopes of enriching their experience. Deep learning will prove very useful in automating the process of extracting audio data, removing the noise in the sample, and simultaneously classifying the music according to predetermined parameters such as tempo, rhythm and timbre through the use of neural networks and training algorithms (See Sarkar et al. (2020). for more information)

2 Illustration

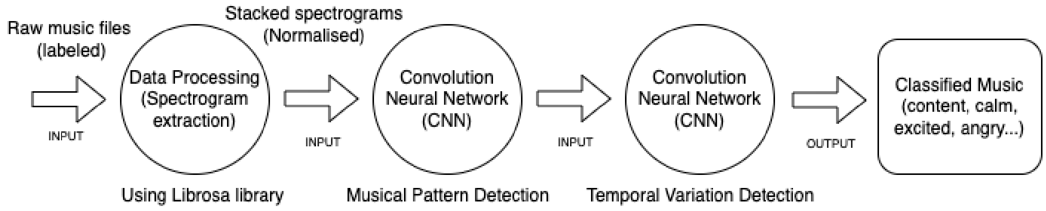

Figure 1: Diagram showing overall proposed music classification model

3 Background and Related Work

We know there are several different aspects of music which evokes certain emotions, ranging from lyrics, rhythm, tempo, pitch. With the use of machine learning, we can relate the impacts of different aspects of music to the sentiment it brings. In the field, this is called Music Emotion Recognition (MER).

3.1 Feature Extraction and Traditional Machine Learning

"A survey of music emotion recognition" Han et al. (2022) explains how training a traditional machine learning model can classify Music Emotion Recognition (MER). This kind of model requires two parts. First for specific features to be extracted and then to be analyzed by an algorithm. Since music is not intuitively quantifiable, researchers have selected certain aspects of music which can be extracted. Features of the track's audio feature (waves, timbre), symbolic feature (pitch, interval, duration). An emotion model must also be chosen so that the sampled music can be sorted into clear categories. The most commonly used by deep learning models to categorize perceived emotions evoked by music is the Russell's Circumplex Model. Emotion is determined by valence (positive/negative emotion) and arousal (passive and activated emotion evoked). Features are extracted using preprocessing methods such as framing, windowing, spectrogram extraction, main track extraction and preprocessing tools such as Pysound, MATLAB, or with python packages like Librosa, pretty_music, music21. The preprocessing results in MFCC, spectrogram, key, BPM, melody and other musical features which are then fed into machine learning levels. They are then fed into machine learning models, and will result in different types of classifications, depending on the model. One type is Song Level Categorical MER whose classification model is support vector machines (SVM) and uses k-nearest neighbors, decision tree, random forests and native Bayes and classifies songs to predefined emotions. Another is Song Level Dimensional MER uses regression such as support vector regression, and gaussian process regression to classify the song into different dimensions (such as valence and arousal). By using these algorithms, the machine is able to classify the emotion based on the musical piece.

3.2 Heirarchical Framework Classification

The paper "Automatic Mood Detection and Tracking of Music Audio Signals" Lu et al. (2006), is an example of how this system works. Musical features such as tempo, loudness, pitch change were automatically extracted through MIDI Files, converted into frames containing spectral bands. Analysis using the Fourier transform Frequency domain converted the images to features such as timbre (based on spectral shape) and intensity (based on subband energy). The information was then fed into a hierarchical framework, which then classified the music by emotion (contentment, depression, exuberance, and anxiousness). By using a hierarchical framework, the music features known to be more impactful in setting the mood were able to be weighed more heavily in musical analysis. As a result, 800 short clips of acoustic, classical music with a testing accuracy of 86.3 percent.

3.3 Deep Learning Neural Networks

The paper "Music emotion recognition using recurrent neural networks and pretrained models" Grekow (2021) utilizes deep learning in training machines to recognize the sentiment that music brings. In contrast with the traditional machine learning method, no specific musical aspects need to be extracted prior to running it through the algorithm. Instead the system will automatically extract suitable features depending on the data. This is important to bring good results, as often the different quality and methods of extraction of different musical pieces can lead to differing results. Through deep learning, convolutional or recurrent neural networks (CNN and RNN) can be used as an end to end processing framework. CNN Models are frequently used, as it can learn feature representations based on data effectively.

3.4 Bi-Modal Deep Boltzmann Machine

Another example of utilizing Deep Learning in MER is outlined in the paper "Bi-Modal Deep Boltzmann Machine Based Musical Emotion Classification" Huang et al. (2016). The Bi-modal Deep Boltzmann Machine architecture (DBM) is a deep neural network which bases its probability on the Boltzmann Distribution and is based on the Restricted Boltzmann Machine. It uses two layers of DBM networks (one for audio, one for lyrics) as well as one additional layer to join the two. Once the machine has selected suitable features of the given track, these features are inputted into SVM, as used in traditional machine learning, and classified into the suitable categories. The result of this proved to outperform other singular modality models, which proves that the DBM model is effective in determining music sentiment due to its consideration of the relationship between lyrics and audio features.

3.5 Visual Geometry Group

Visual Geometry Group Net, an improved version of CNN was used in paper "Recognition of emotion in music based on deep convolutional neural network"(Sarkar et al., 2020) to explore the performance of different audio features to the emotion it evokes. Due to the increased number of layers, music recognition accuracy was increased relative to CNN models however the model struggled to identify arousal and time series nature of audio could not be properly represented in the model (See Sarkar et al. (2020) for more information)

Music is not intuitively quantitative. The process of quantifying and identifying musical features is extremely important and can skew results if done inconsistently. With deep learning, models can consistently and automatically extract musical features, increase their accuracy with larger sample sizes, and can create multi-dimensional analysis which considers audio features, musical symbols and lyrics altogether.

4 Data Processing

In the domain of acoustic analysis, there are a multitude of options regarding the format of the input data for a Machine Learning Model. Such options vary in complexity, computational burden, the scope of parameters, and precision of parameters/features, amongst other components. Therefore, one must critically determine how, systematically, the dataset will be processed.

In consideration of the aforementioned concerns, we have decided to process data into four unique features, each capturing a different fundamental characteristic of a song. The paradigm we have agreed on, stemming for the paper "A survey of music emotion recognition" [3], is that the fingerprint of a song can be described by four key musical dimensions: tonality, harmony, timbre, texture and rhythm. Thus, these four features provide a reasonable level of complexity while still capturing fundamental dimensions of music closely correlated with genre as well as the emotions it is likely to evoke. The features are the following:

1. **MFCCs (Mel-Frequency Cepstral Coefficients):** These coefficients capture the short-term power spectrum of sound, providing information on the variety of timbre texture in the song, and distinguishing between different tonal qualities.

2. **Spectral Contrast:** This extracts the difference between peaks and valleys in a sound spectrum, aiding in the differentiation between timbral tones; it offers clues about the harmonic and non-harmonic content of a song.
3. **Chroma:** This feature provides a snapshot of the energy distribution across twelve pitch classes, corresponding to the twelve musical notes in an octave. It is pivotal for understanding harmony, chord progressions and tonal structures.
4. **Tempogram:** Tempograms provides a representation of rhythm and tempo fluctuations, by capturing the onset strength envelope over time. This feature reveals intricate rhythmic patterns and variations in tempo.

We intend on utilizing the Librosa Python Library, a powerful tool for audio and music analysis to process raw (labeled) audio files in order to extract the features listed above.

5 Architecture

The proposed architecture will be a hybrid network consisting of both CNN and RNN (See Yu (2021) for more information). CNNs are great for pattern detection from spectrograms which is precisely the input format we will gather in our data-processing stage. Meanwhile RNNs can capture temporal dependencies in the audio sequences. The idea is to use the CNN layers for extracting spatial features from audio spectrograms, and then feed the output of these layers into RNN layers to capture the temporal dependencies and variations inherent in music data. In terms of the heterogeneity of the parameters, we have adopted a feature stacking method, where the data is normalized and concatenated so that no feature initially overpowers the other. This will ensure the input data is a hierarchical classification scheme that could be implemented to improve classification accuracy by breaking down the classification task into simpler, more manageable tasks, which could lead to more accurate and interpretable classification outcomes.

6 Baseline Model

As our baseline model, we selected the Support Vector Machine (SVM) which is a supervised learning algorithm, commonly used for classification tasks. It attempts to find a hyperplane that best divides a dataset into classes. Given the high-dimensional nature of audio data, SVMs can be particularly useful as they can excel in handling high-dimensional data. For this task of music emotion classification, we will use the earlier mentioned features to find the best fit non-linear hyperplane and evaluate our accuracy and F1-score.

7 Ethical Considerations

Music is a creative art form. As a result, many music tracks are copyrighted and proper usage rights and licensing must be obtained in order to be used for machine learning legally. To avoid violating usage rights and copyright law, the team has agreed to use existing, royalty free and public domain music for our testing and training data. This ensures proper licensing which allows us to use the music for research purposes.

As a result, the team acknowledges there is potentially some bias which comes with using royalty free and public domain music from existing sample sets. Different kinds of music exist in all cultures. Since we will be collecting sample sets from royalty free and public domain websites with English writing, our dataset may have more samples with music reflective of English speaking cultures. Furthermore, music is a subjective art form. Different cultures have different interpretations of music. By having a limited data sample, our network will likely be trained on mainly public domain datasets created from English speaking cultures. By having this limitation, our data could potentially misidentify emotions music brings because it lacks cultural contexts. This could result in reinforcing stereotypes of certain cultures and reinforcing inequalities of misrepresentation and under-representation in the music industry.

8 Project Plan

As a team, we share a common goal for the project to be a source of pride for us, something worth showcasing on our future portfolios. In order to achieve this goal, we have agreed that as a team, we need an established communication platform, delegate clear responsibilities, and ensure efficient collaborative work is being completed.

The process for designing the model should be shared amongst the four group members, however certain members will be responsible for ensuring specific tasks are completed by the team on time. Additionally, the write up portion will be completed individually amongst team members and reviewed by the team to agree the correct and accurate information is captured.

8.1 Communication

We have agreed to all communicate on our Discord channel regarding the project. When in need of a team member, we will first ping them. If we are unable to get a hold of them, we will phone call them as a last resort. We have all shared our phone numbers on Discord.

8.2 Responsibility Delegation

Before beginning each deliverable, the team will have a kick off meeting to go over the assigned tasks, highlight key ideas they would like to include, and re-evaluate if the assigned project responsibilities from the assigned task table (Table 1.0, 2.0, 3.0, 4.0) is still suitable for team members. In general, idea generation will be done as a group. To ensure all project requirements are met, team members are assigned to lead certain tasks, to ensure they are completed by the team and properly documented in the report. Writing tasks will be drafted individually and reviewed by all members of the group to ensure accuracy.

Table 1 outlines the tasks team members have assigned and completed during the Project Proposal.

Table 1: Project Proposal Plan Task Assignment

LEAD	ASSIGNED TASK	DUE
All Team Members	Group Brainstorming Session - Each team member presents interesting potential project ideas	Oct 11
Jennifer	Writing Tasks: Background and Related Work, Ethical Considerations, Project Plan, References, Report Formatting	Oct 11
Chielotam	Writing Tasks: Introduction, Illustration, Risk Register, Background and Related Works	Oct 11
Daniel	Writing Tasks: Architecture, Baseline, Risk Register, References, Abstract, Report Formatting	Oct 11
Jonas	Writing TasksData Process, Architecture, Baseline, Report Formatting	Oct 11

In Table 2, the process for designing the model should be shared amongst the four group members, however certain members will be responsible for ensuring specific tasks are completed by the team on time. Additionally, the write up portion will be completed individually amongst team members and reviewed by the team to agree the correct and accurate information is captured.

Table 2: Project Progress Report Task Assignment

LEAD	ASSIGNED TASK	DUE
All Team	White board Idea coding session	Oct 20
Jennifer	Obtaining data (ensure data is copyright free)	Oct 28
	Collection dataset of at least 1000 samples	Oct 28
Chielotam	Producing 1 Qualitative and Quantitative result	Oct 28
	Feasibility Considerations throughout entire modeling process	Oct 28
	Writing Task: Brief Project Description	Oct 30
Jonas	Writing Task: Noteable Contribtutions - Data Processing, Primary Model	Oct 30
Daniel	Writing Task: Noteable Contribtutions - Baseline Model, Primary Model	Oct 30
All Teams	Project finalization and overview	Nov 2
	citations	Nov 2

Table 3 reflects the current decision on how the project presentation will be divided. The current plan is to divide it into video segments with each team member responsible for writing and filming their own assigned segment. Team members will look over eachother's written script and give feedback before recording. Then the video will be edited together and submitted. Depending on available time and resources, crentain aspects of the video may change. Teams have agreed to complete a rough draft by November 17th while a final copy is due November 26th.

Table 3: Project Presentation Task Assignment

LEAD	ASSIGNED TASK	DUE
All team	Further development of model and training?	Nov 9
Chielotam	Problem, Takeaways/outro/citations	Nov 17
Jennifer	Data Collection, Demonstration	Nov 17
Jonas and Daniel	Data processing, Quanitative and Qualitative Results:	Nov 26
Jonas	video editing and presentation flow	Nov 26

Table 4 reflects the plan for the final report. Team members will plan have completed the main points of the report and a rough draft by 17th of November and final copies ready on November 30th.

Table 4: Project Final Report Task Assignment

LEAD	ASSIGNED TASK	DUE
Chielotam	Introduction, illustration, Project Quality, Last Edit and structure, grammar, citations	Nov 30
Jennifer	Background and Related Work, Discussion, Ethical Considerations	Nov 30
Jonas	Data Processing, Qualitative Results, Architecture	Nov 30
Daniel	Architecture, Baseline Model, Quantitative Results	Nov 30

To ensure efficient collaborative work, and prevent writing over each other's code, the team has agreed to use Google Collaborate, as linked in section 10. Additionally, team members will send a message on the discord chat when they are planning to work on the code.

In order to function well as a team, mutual accountability and respect for each other is essential. Team members should try their best to reach the internal deadlines while delivering quality work. We also recognize the importance of being flexible and empathetic, as we are students juggling a heavy course load and other various commitments.

9 Risk Registry

As with all projects, regardless of the level of complexity, considerable risks might come into play that might delay the project or lead to a poorer quality result. For this project, we have identified four possible major risks and subsequent mitigation strategies. The likelihood of these risks have also been ranked on a scale of 1 - 5 (5 being the most likely). They are listed as follows:

9.1 The subjectivity of music and emotions

Likelihood: 3

Concern: This is a major risk because music often contains many intricate details like tempo, melody, and harmony (audio data) which could contrast with other elements like the lyrics used. It could also be hard to classify due to different perceptions of emotions per certain groups of people.

Mitigation strategy: A recommended approach is to reference existing research on the topic of music emotion analysis.This emphasizes the need for considering a variety of different sample data.

9.2 Copyright infringement

Likelihood: 2

Concern: It is possible that the sample data we choose to analyze might not be regulated for use. This would stem from choosing data from unreliable sources.

Mitigation strategy: The solution to this would be to choose music that has been marked or identified as non-copyrighted and/or royalty free music.

9.3 Team member dropping course

Likelihood: 3

Concern: If a team member drops the course due to unforeseen events, we may not be able to finish the project on time. The likelihood of this event is uncertain but critical as it is difficult to know every team member's personal circumstances at all times.

Mitigation strategy: The potential solution could be simplification of the project to which the current members could handle and implement a divide-and-conquer approach.

9.4 Time to train model

Likelihood: 4

Concern: This may cause the team to submit an unfinished project which is a major risk.

Mitigation strategy: A basic yet functional model is always preferable to an unfinished complex one. Therefore, we can be smart in our choice of models. i.e., implementing an existing model and adding something simple yet novel.

10 Google Collab

https://colab.research.google.com/drive/1EhqqafIbHAustkjzFOarG6DIH-zbJPJG?usp=sharin

REFERENCES

J. Grekow. Music emotion recognition using recurrent neural networks and pretrained models. *Journal of Intelligent Information Systems*, 57(3):531–546, 2021.

Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6):16, 2022.

Moyuan Huang, Wenge Rong, Tom Arjannikov, Nan Jiang, and Zhang Xiong. Bi-modal deep boltzmann machine based musical emotion classification. In *Artificial Neural Networks and Machine Learning – ICANN 2016*, volume 9887, pp. 199–207. Springer, 2016.

Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):5–19, 2006.

R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha. Recognition of emotion in music based on deep convolutional neural networks. *Multimedia Tools and Applications*, 79(1-2):765–783, 2020.

Y. Yu. Research on music emotion classification based on cnn-lstm network. *IEEE Access*, 2021.