EmoSense: Music Classification Based on Audio, Video and Lyrical Content

Srishti Singh
srishti20409@iiitd.ac.in

Shreya Bhatia
shreya20542@iiitd.ac.in

Tarushi Gandhi
tarushi20579@iiitd.ac.in

Ashwin Sheoran
ashwin20288@iiitd.ac.in

Ashish Kamathi
ashish20364@iiitd.ac.in

Vedant Patil
vedant20348@iiitd.ac.in

Abstract

Music classification is a challenging task due to the subjectivity, complexity, a lack of labelled data, constantly evolving music, cross-genre songs, and cultural differences . These issues can make it difficult to develop a comprehensive and accurate music classification system capable of classifying all types of music, especially given the ever-changing nature of music and the influence of cultural differences. Furthermore, the subjective nature of music classification, as well as the presence of cross-genre songs, can lead to classification inaccuracies. We intend to address these issues individually while developing an effective classification model.

Keywords— I nformation Retrieval , Music Mood extraction , Sentiment Analysis, Recommend-er System

1. Motivation

This project offers a fascinating chance to investigate the potential of artificial intelligence and machine learning to categorise music based on lyrical, audio, and video content. By working on this project, we can learn more about how various listeners perceive and categorise music. This project will make use of cutting-edge Machine Learning methods to accurately categorise music into different emotions expressed based on its audio, video, and lyrical content. Finally, using machine learning, this project will enable us to more fully comprehend the beauty of music.

2. Related Work

2.1. Multi-modal Music Emotion Classification based on audio and lyrics [2]

This paper uses multi-modal fusion emotion classification method based on audio and lyrics. Lyrics have been classified using Bert model, then LFSM based equalization was performed on the lyrics emotion classification results using the sentiment dictionary. For features in audio data they used Mel Frequency Cepstrum Coefficient, spectrum centroid and frequency-based energy distribution which are fed into LSTM model for music emotion classification. Their new fusion method achieved 5.77 percent and 4.03 percent improvement over the linear weighted multimodal and LASM fusion techniques.

2.2. Based on Improved Convolutional Neural Network [1]

The study describes a method for music emotion recognition that combines mel-frequency cepstral coefficient (MFCC) and residual phase (RP) to extract low-level audio features. Convolutional recurrent neural network (CRNN) is used to extract time-domain, frequency-domain, and sequence features of audio. Bidirectional long short-term memory (Bi-LSTM) network is used to obtain sequence information of audio features. The features are then fused and input into a softmax classification function with center loss function to recognize four music emotions. The proposed method achieved 92.06 percent accuracy, outperforming other methods. The method provides a new approach

2.3. Audio-Visual Approach to Music Genre Classification through Affective Color Features [5]

This paper adopted a methodology based on extracting colour and hue information from music videos. From the audio data they used pleasure, arousal and dominance. They achieved an accuracy of 50.13 percent using only visual data with SVM. Their findings revealed that while accuracy is negatively impacted by spectral and timbral characteristics, it is positively impacted by the mixture of visual data with chroma and rhythm descriptors.

2.4. Deep learning-based late fusion of multimodal information for emotion classification of music video [4]

The article discusses the creation of a diverse music video emotion dataset and its use in testing four unimodal and four multimodal convolutional neural networks (CNNs). The best unimodal classifier is integrated with corresponding music and video network features to create a multimodal classifier, which integrates whole music video features and uses a SoftMax classifier for final classification using a late feature fusion strategy. The multimodal structure achieved an accuracy of 88.56 percent, an f1-score of 0.88, and an AUC score of 0.987, demonstrating better performance than each unimodal emotion classifier.

2.5. Music video emotion classification using slow{fast audio{video network and unsupervised feature representation [3]

A multimodal architecture with an audio-video information exchange and boosting mechanism was used to process the music and video data. End-to-end training was performed on a number of supervised and unsupervised networks, and several evaluation criteria were used to assess the performance. They trained an autoencoder network, with the encoder utilizing two multimodal structures and a large number of convolutions. Utilizing 2D/3D convolution, the decoder network was created to be compact and using less computational resources. With the least amount of computing effort, their top classifier achieved 77 percent accuracy, an area under the curve score of 0.94 and a f1-score of 0.77.

3. Novelty

Looking at the work done so far in the field of music classification, most of them focus on analysing audio and lyrics to predict music genre or emotion of the video. Putting music into one category limits accountability of genre fusions, range of emotions, variety of styles, historical contexts and cultural influences. Therefore we aim to incorporate analysis of Video, audio and lyrical data extracted from a music video to more accurately label the distribution of emotions in it. There has been no work up to date that uses all three forms of data to classify the emotions of a music video.

4. Acknowledgement

We are thankful to our instructor Dr. Rajiv Ratn Shah and the Teaching Assistants of our course Information Retrieval for guiding us through this project area and proposal.

5. Project Plan

The individual task have been distributed in following manner -

Data Collection	Ashwin, Srishti, Vedant
Data Visualisation , Pre Processing	Ashish, Shreya , Tarushi
Model Building	Tarushi, Ashwin, Shreya, Srishti
Analysis and Performance	Vedant, Ashish
Interface Development	Ashish, Vedant, Srishti, Shreya
Evaluation and Feedback	Ashwin, Tarushi
Report Writing	Srishti, Ashwin, Shreya, Tarushi, Vedant, Ashish

References

[1] Xiaosong Jia. A Music Emotion Classification Model Based on the Improved Convolutional Neural Network. *Computational Intelligence and Neuroscience*, 2022:6749622, 2022. 1

[2] Gaojun Liu and Zhiyuan Tan. Research on Multi-modal Music Emotion Classification Based on Audio and Lyirc. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 2331–2335, June 2020. 1

[3] Yagya Raj Pandeya, Bhuwan Bhattarai, and Joonwhoan Lee. Music video emotion classification using slow–fast audio–video network and unsupervised feature representation. *Scientific Reports*, 11(1):19834, Oct. 2021. Number: 1 Publisher: Nature Publishing Group. 2

[4] Yagya Raj Pandeya and Joonwhoan Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2):2887–2905, Jan. 2021. 2

[5] Alexander Schindler and Andreas Rauber. An Audio-Visual Approach to Music Genre Classification through Affective Color Features. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 61–67, Cham, 2015. Springer International Publishing. 1