Music Emotion Classification Experiment by using Deep Learning method

Tri-Nhan Do
Advanced Program in Computer Science
Faculty of Information Technology
University of Science, VNU-HCM
dtnhan@apcs.vn

Tri M. Nguyen
Advanced Program in Computer Science
Faculty of Information Technology
University of Science, VNU-HCM
nmtri17@apcs.vn

***Abstract*—Music mood classification is one of the most important features in music recommendation. A song include lyrics and audio, we try to classify each of them. To enhance the quality of music classification, the authors try to do many Deep Learning methods relate to both lyrics and audio. Those methods we try to use include CNN, LSTM for lyrics classification and ResNet, LSTM for audio classification to recognizing the mood. The authors use two kind of datasets are MoodyLyricsPN for lyrics classification and MTG-Jamendo 2019 for audio classification. The authors manage to achieve the accuracy of 72.4% with CNN model, 66.6% with LSTM model. The results are not that impressive, but this experiment leads to many approaches for future work when we have not yet applied the Hyperparameter tuning methods and dataset we use for training smaller than the baseline.**

I. Introduction

The number of people listening to online music websites is increasing dramatically accompanied with services to improve user experiences. One of the most significant services is music recommendation system. These recommendation systems filter information to predict users' preference of a certain song.

There are two main approaches for music recommendation systems, which are Collaborative Filtering and Content Based Filtering [1]. Collaborative Filtering approach bases on similarity between users' behaviours, activities or preferences. The major problem of this approach is that it requires a large dataset to achieve high accuracy, which can lead to a "cold start" problem. In contrast, Content Based Filtering is the most common techniques for music recommendation system. In this approach, the model closely analyzes the characteristics of a song to make a recommendation. This leads to the demand for classifying a song.

According to proposed methods for music classification, music can be classified according to artist, genre, instrument or mood of the song [2]. However, classification based on songs' mood is the most interesting and challenging approach because mood is hard to define and it depends on different people. [3] The approach to classify music varies in different ways. One of a potential one is to use subjective human feedbacks or user tags [4]. However, a drawback of this method is that it is hard to collect datasets since users are not always willing to provide their feedbacks. Of several hundred users surveyed, listeners indicated that vocals (29.7%), lyrics (55.6%), or both (16.1%) are among the most salient attributes that they notice in music [5]. Another approach is to classify based on its characteristic like audio and lyrics [6] [7] [8] [9]. However, not only does the song's characteristic but also feeling of different people based on the situations that it is played makes it difficult to evaluate the mood of the song correctly.

There are several researches on this topic and standard ground truth for it has been improved day by day such as MIREX [10]. Although dataset collected by MIREX has high reputation and is created by many experts, it is not enough for training in deep learning. MoodyLyricsPN is another lyrics dataset that contain about 5000 songs's lyrics that will help the author to do song classification in lyrics. On the other hand, MTG-Jamendo datasets will be used for song classification in audio. Both datasets are reliable, espically the MTG-Jamendo dataset. [11].

To determine mood of a song, some studies use lexicon based, whereas others apply machine learning approaches [12] [13] [14]. Our proposed methods are to combine both of the two methods, using NLP for feature pre-processing, combine lexicon's confidence and word embedding weight for featuring and using Convolutional Neural Network, Recurrent Neural Network model for lyric and using ResNet, Keras model for audio. The authors try to train with one-hot encoding and experiment some pre-trained word vector extraction methods like fastText, word2vec, GloVe to choose the most proper feature extraction for lyric features. On the other hand, the ResNet model for audio recognition is pretty complete in Dcase dataset and we apply the code for Jamendo dataset to check whether it suitable when change from 10 class to 56 class. We also fit LSTM model with the input of Mel spectrogram, the model is still run for the result.

The rest of the paper is organized as follow: Section 2 reveals some related works on music mood classification, section 3 introduces the datasets and the authors' proposed methods using audios and lyrics analysis, section 4 talk about the author's experiments, section 5 will reveal the result and finally, the author will draw a conclusion and discusses more future works in section 6.

II. Related work

The approach of the authors is to analyze mood of a song based on its lyrics and audios. Therefore, the problem of sentiment analysis from textual and sound analysis are the main focus of this paper. There are many works that were conducted in sentiment analysis field. All the of related work method use machine learning, some bases on lyrics, some bases on audio and some bases on both. Some of the most significant related works are listed below. The first two ones include method relate to lyrics analysis, the next two ones introduce to you method relate to audio analysis and the last introduce the method that use both lyrics and audio.

A. Music mood classification using machine learning

A recommendation system was built based on a Naive Bayes classifier [15]. By analyzing lyrics of songs, the method classifies training dataset containing 1000 songs into 2 moods, happy and sad. The most significant contribution of this project is the creation of dataset which was filtered, labelled, and publicized. Moreover, the Naive Bayes in this project yields the result of 72.5%.

B. Word Vector for Sentiment Analysis

This project proposes the approach of combining unsupervised and supervised method for capturing between semantic and sentiment similarities among words. The comparison between the proposed approach of word representation learning model and some other commonly used vector space models such as Latent Semantic Analysis, Latent Dirichlet Allocation, and Weighting Variants indicates that when capturing word representation instead of latent topics, the performance of the model improves.

C. The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification

The paper was submitted to DCASE 2019 for Acoustic Scene Classification task. In this paper, those CNN method like Resnet, Desnet and VGG was proposed for acoustic scene classification. This paper also show that very small or very large RFs can cause performance degradation, but deep models can be made to generalize well by carefully choosing an appropriate RF size within a certain range [16]. The author will use the ResNet method which was proposed in this paper to try with music mood classification.

D. CNN based music emotion classification

The paper introduce the method of using CNN on song's audio to find song mood/emotion. This project using music's spectrograms (which is pretty similar to the authors's) that contain both original time and frequency domain infomation, which are contained on the standard CAL500 and CAL500exp dataset. Based on their claim, the result out-perform the state of the art method. [17]

E. Music Mood Detection Based On Audio And Lyrics With Deep Neural Net

The proposed method use both lyrics and audio, which was reproduced the implementation of traditional feature engineering based approaches and propose a new model based on deep learning [18]. The method use CNN for lyrics only and audio only. It reuse the unimodel architecture for fusion (both music and audio). This method give their author conclustion that Deep Learning method work well than traditional method.

III. Proposed method

In this section the authors will introduce all the method that we use to classify songs's audio and lyric (but not both). All method use are Deep Learning method. We reproduce the code to make CNN and LSTM for lyrics classification. For the audio, the authors try to reuse the ResNet model implemented in Pytorch which was mentioned above and the LSTM model the authors build on Keras.

A. Lyric:

1) Dataset: The dataset we will use is the MoodyLyricsPN dataset which was introduced to contain 5000 songs divided into two label positive and negative with ratio 1:1.

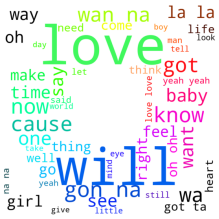

Fig. 1. WordCloud of Positive class MoodyLyricPN dataset

Fig. 2. WordCloud of Negative class MoodyLyricPN dataset

2) CNN: With CNN model we deploy4 layers: 1 embedding layers, 1 convolutional layer and 2 fully connected layer. For the convolutional layer, we use the length 3 of the 1D convolution window. All hidden layers are equipped with the rectification (RELUs) non-linearity. We use spatial drop out 1D after the embedding layer and dropout after the dense layer but not after the convolutional layer. From the song lyrics, we convert it to sequences of tokens and pad it to ensure equal

length vectors of 250. Then we put it into the embedding layer before extract feature with convolutional layer because vectors is high dimensional and sparse. After the convolutional layer, we use global max pooling layer to feed directly feature maps into feature vectors. From feature vectors, we apply fully connected and sigmoid function to calculate probability that this song will have a positive or negative mood. Between this, we have set dropout rate as 0.25 for the dense layer. To rate our model, we use binary-cross entropy function as an objective function and Adam optimization algorithm to optimize our parameters.

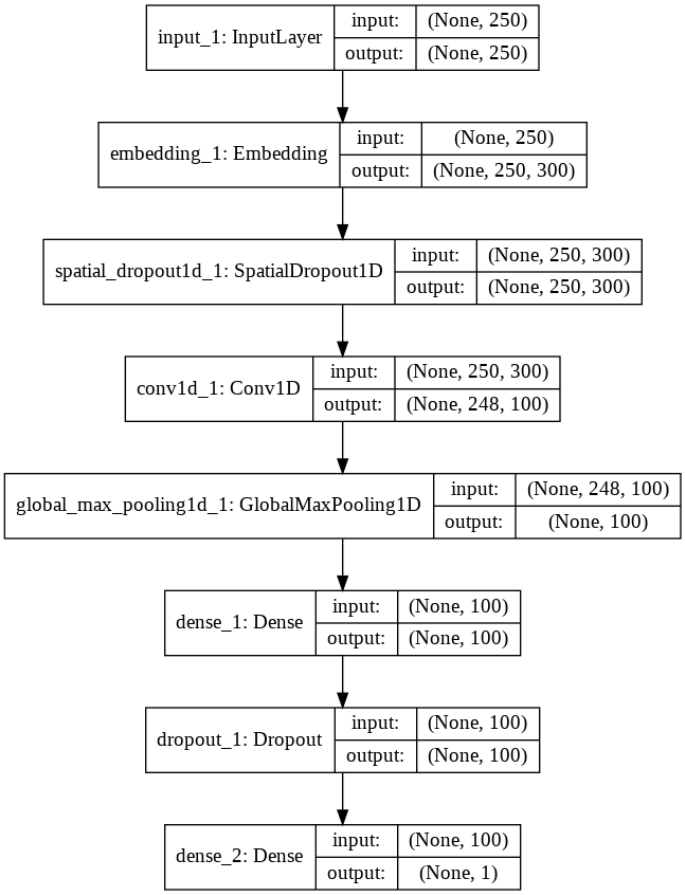

Fig. 3. CNN model for Lyric

3) LSTM: With LSTM model, compare to CNN, this RNN model learns information from immediate previous step it can be updated while training the neural network. The embedding layer encodes the input sequence into a sequence of dense vectors of dimension, the LSTM transforms the vector sequence into a single vector, containing informa- tion about the entire sequence. The other hyper parameters like dropout, batch size are similar to that of CNN

B. Audio:

1) Dataset: We use the MTG-Jamendo dataset which was provided in "the MediaEval 2019 Emotion and Theme Recognition in Music" challenge. MediaEval is a multitask contest

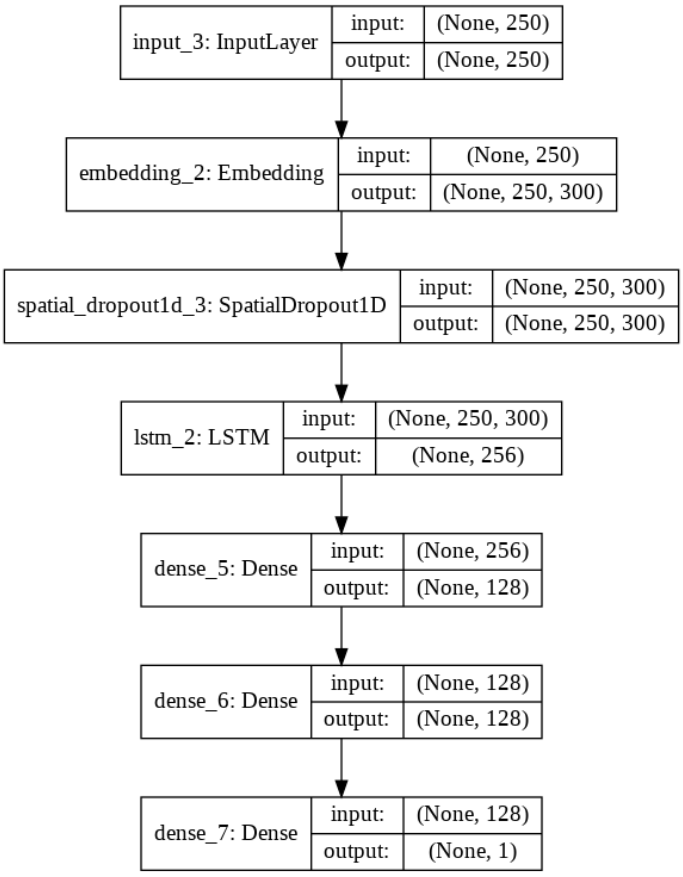

Fig. 4. LSTM model for Lyric

and in 2019, they provide a large amount of datasets in MER (music emotion recoginzation) task, which about 18,486 songs's audio with 56 mood/theme tags.

Initially we planned to combine audio and lyrics, so we emailed jamendo to ask for the API, but when we ran, we only received 1500 song have lyric, 1/12 of dataset, so we decided Experiment on audio completely

The author download the mel-spectrogram only dataset which was provied by some custom script and try to apply to the ResNet and LSTM model.

Mel-spectrogram is the result of collections of actions like taking the sound input, separate it to window, generate mel scale and generate the spectrogram. Mel-spectrogram can be saw as better representation of sound that can be heard by people, which is suitable for task like speech recognition or music recognition.

2) ResNet: We use the model metioned in [16] since the model use in Acoustic scene classification task, which pretty similar to mood/theme recoginize task, so the author try to test this model on the MTG-Jamendo dataset. The model is ResNet-28 with small Receptive Field. It has been prove in its paper that small Receptive Field will prevent model from overfiting and gain higher result. Beside, the model use Adam optimizer and Mixup data augmentation.

3) LSTM: We create a model which contain 4 layer, the first layer is Input layer which receive the input of List of MelSpectrogram, them we dropout with rate of 0.3. The next layer is LSTM which output an array of 256 for each sample, then we go through 2 dense layer, then take activation funtion softmax for multi-class classification problem, we get output is 56, corresponding to the number of labels, the loss function when compile this model is categorical crossentropy, optimize with Adam with learning rate of 0.001.

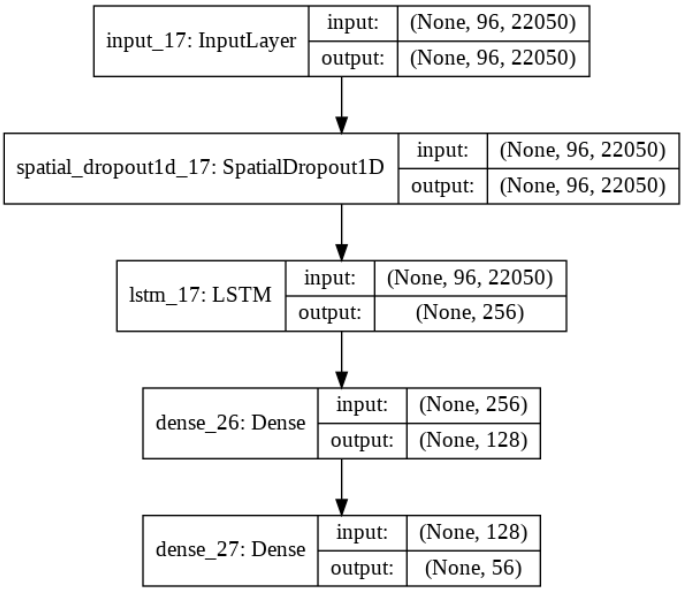

Fig. 5. LSTM model for Audio

IV. Experiment:

A. Environment setup:

1) Lyrics: In the lyrics task the author set up Colab environment with python 3.7 with all necessary package.

2) Audio: The author follow the instructions to set up the model, that is installing, activating conda environment and other necessary library. We set it up so it can run and support by GPU. Our GPU has the Nvidia Guadro K6000 card which can be used for CUDA support. However, for ResNet model, since the recommended PyTorch version can not apply to GPU with compatibility 3.5, the author must install PyTorch from source which cost a lot of time. The authors also must downgrade the librosa version to 6.3.0 to be compatible with the model and configs suitable path to our datasets in the config file for ResNet models.

B. Pre-processing:

1) Lyrics: Before employing machine learning approach to extract sentiments, the typical pre-processing procedure is applied. For both CNN and LSTM model, they have the same steps.

First we must remove all lyrics that has length too short or instrumental music. Those are about 400 lyrics of these type, which reduce the number of sample from 5000 to 4600. Although we lost about 8% of total number of samples, we believe the remain amount still big enough to use.

After lowercase all the remain lyrics, next we must remove all chorus and versus from songs lyrics. Those are prefix of sentence that tell us that how many time that sentence is told, for example: ”chorus(x3) Ha” mean we must repeat Ha three times. Although these thing can be important, we still remove them since they are not impact much to our model.

We then apply text normalization techniques, which is lemmatization from the NLTK library to achieve the root forms of inflected words. We choose WordNet Lemmatizer that uses the WordNet Database to lookup lemmas of words. Also, accroding to Erion Çano, we should remove stopword {“the”, “these”, “those”, “this”, “of”, “at”, “that”, “a”, “for”, “an”, “as”, “by”} so we do the same thing.

We use LabelEncoder to normalize labels and transform non-numerical labels Negative and Positive (as long as they are hashable and comparable) to numerical labels (encode labels with value 0 and 1 for two above class)

Finally, the dataset is divided by train, validation and test share 70:10:20. However, the number of lyrics crawled was only 4600 songs, because some songs need copyright. To maintain fairness when comparing models, we kept 1,000 lyrics for evaluation (equal to the number of lyrics the dataset author author used for testing), so we had 3150 for training and 450 for validation in deep neutron network models. The result from the Test will be compare with the baseline.

2) Audio: The audio pre-processing are different for two kinds of models.

ResNet:

Since the input of ResNet model are audios, we must change the pre-processing step in the model.

The author must change the mel-spectrogram input to suitable with the mel-spectrogram standard of model. By padding and cutting the mel-spectrogram to fit the model layer shape, the author finally gain the input which can be run on the model.

LSTM:

With each audio sample we receive 2D matrix Mel spectrogram with time and Mel feature, the value is Sound Intensity dB. The duration of each audio is different, therefore we use zero-padding to get the same size (we are also going to try reflection padding), the shape return of each spectrogram is (96,22050) with 96 is number of Mel feature.

C. Model running:

1) Lyrics: Here is a full information about how the authors create embedding layer with pre-trained fastText. The author choose the fastText with 1 million words trained on Wikipedia in 2017, UMBC web base corpus and statmt.org news dataset. Each word returns a vector with 300 elements, which means that word is represented in a spatial coordinate with 300 dimensions. From 21391 words from 3150 songs of the train dataset is embedded with the corresponding word vector in fastText to build up 21391 300-dimension vectors. Then, each word in a song is replaced by its corresponding vector.

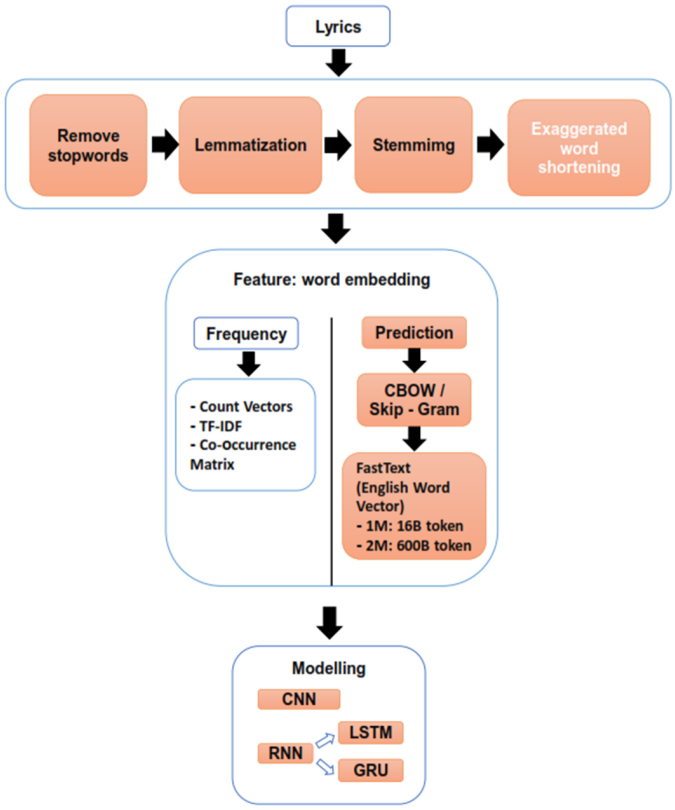

Fig. 6. RNN - LSTM model

2) Audio: Resnet with input

Although we can run successfully with the DCASE2018 Acoustic Scene Classification Dataset, we meet trouble when running a large amount of data samples in MTG-Jamendo dataset. The pre-processing takes much time and memory, which our GPU can't meet the require. Therefore, we stop after one epochs.

LSTM with input Mel spectrogram

Because we have total 56 classes, we make one hot encoding for each of song input, therefore the shape of label is (number of audio, 56) We put whole Mel spectrogram to LSTM model with the rate of dropout layer is 0.1 to avoid early overfiting, the number of epoch we choose is 30. The optimize Adam is use for this process with learning rate 0.001.

V. Result:

A. Lyrics:

After training model CNN and LSTM, we evaluate the test set with 1000 lyric song of MoodyLyricPN dataset

TABLE I
OUR METHOD INITIAL RESULT

Mlpn	CNN	LSTM
75.63%	72.4%	66.6%

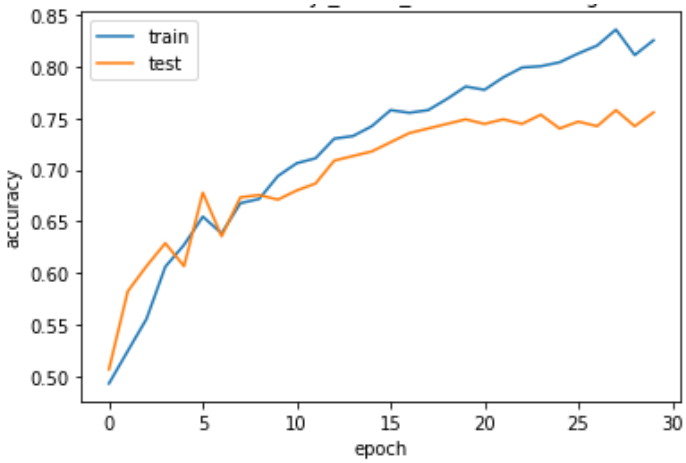

Fig. 7. CNN training result

B. Audio:

The result achieved with audio classification is quite bad, the ResNet model we clone from github running on dataset Dcase returns 79.5% result, but when we fit it into the Jamendo dataset with 5 times the number of classes, the result decreases pretty much, at the first epoch the accuracy for validation result was 0.4%, so we delayed this approach.

With LSTM mehod we built the model ourselves, currently waiting for the results to be returned. We will then convert the results into ROC measurements to compare with the results of MediaEval's competition.

VI. Conclusion:

In this paper the authors have introduce to many methods relate to both Audio and Lyric. Although Lyric-only model tend to work better, Audio-only model can be improved by running more epochs and having better pre-processing method.

As for the future work, the authors will find a way to improve the overalls performance of ResNet model on the MTG-Jamendo dataset since it still have potential in Music Mood Recognition. the authors also will apply more high-level features for other models. Beside that, the authors will continue the experiment on those model that apply both Audio and Lyric.

Acknowledgment

This work was also supported by Dr.Thao for her guidances and for giving us more time to run the model.

References

[1] Christine Bauer, Marta Kholodylo, and Christine Strauss. Music recommender systems: Challenges and opportunities for non-superstar artists. 06 2017.

[2] Aziz Nasridinov and Young-Ho Park. A study on music genre recognition and classification techniques. In *MUE 2014*, 2014.

[3] Doris Baum. Emomusic - classifying music according to emotion. 01 2006.

[4] Stefan Brecheisen, Hans-Peter Kriegel, Peter Kunath, Alexey Pryakhin, and Florian Vorberger. Muscle: Music classification engine with user feedback. In *EDBT*, 2006.

[5] Andrew Demetriou, Andreas Jansson, Aparna Kumar, and Rachel M. Bittner. Vocals in music matter: the relevance of vocals in the minds of listeners. In *ISMIR*, 2018.

[6] Pasi Saari, Tuomas Eerola, and Olivier Lartillot. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19:1802 – 1812, 09 2011.

[7] Tuomas Eerola and Jonna K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[8] Tuomas Eerola, Olivier Lartillot, and Petri Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. pages 621–626, 01 2009.

[9] C Laurier, Olivier Lartillot, Tuomas Eerola, and Petri Toiviainen. Exploring relationships between audio features and emotion in music. pages 260–264, 01 2009.

[10] Fika Hastarita Rachman, Riyanarto Sarno, and Chastine Fatichah. Music emotion classification based on lyrics-audio using corpus based emotion. *International Journal of Electrical and Computer Engineering (IJECE)*, 8:1720, 06 2018.

[11] Dan Yang and Won-Sook Lee. Music emotion identification from lyrics. pages 624 – 629, 01 2010.

[12] M.D. Devika, Sunitha C, and Amal Ganesh. Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science*, 87:44–49, 12 2016.

[13] Z. Hailong, G. Wenyan, and J. Bo. Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th Web Information System and Application Conference*, pages 262–265, Sep. 2014.

[14] Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. Practical text classification with large pre-trained language models. *CoRR*, abs/1812.01207, 2018.

[15] Sebastian Raschka. Musicmood: Predicting the mood of music from song lyrics using machine learning. 11 2016.

[16] Khaled Koutini, Hamid Eghbal-zadeh, Matthias Dorfer, and Gerhard Widmer. The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification. pages 1–5, 09 2019.

[17] Xin Liu, Qingcai Chen, Xiangping Wu, Yan Liu, and Yang Liu. Cnn based music emotion classification. 04 2017.

[18] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music mood detection based on audio and lyrics with deep neural net. 09 2018.