INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
ALLAHABAD

5^{TH} SEMESTER PROJECT

Text to Song Generation Using Deep Learning

Submitted By:
Ishan Arora (IIT2017501)
Vikrant Singh (IIT2017502)
Srikar Anand (IIT2017504)
Akshay Gupta (IIT2017505)
Naman Deept (IIT2017507)

Submitted To:
Dr. Pavan Chakraborty

CERTIFICATE FROM SUPERVISOR

We hereby declare that the work presented in this end semester project report of B.Tech (IT) 5th Semester entitled Text to song generation using deep learning, submitted by us at Indian Institute of Information Technology, Allahabad, is an authenticated record of our original work carried out from August 2019 to November 2019 under the guidance of Dr. Pavan Chakraborty.

Due acknowledgements have been made in the text to all other material used. The project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Date : 04-12-2019
Place : IIIT Allahabad

Supervisor:
Dr Pavan Chakraborty

Abstract

This report describes the 5th semester project our group is working on, titled Generating music from text and then synthesizing it to a song . The study aims to compose a song based on a user given lyrics. The first step that is required is emotion detection, i.e., detecting mood of a given lyrics, based on which a tune can be composed.

Once the emotion is detected from the given lyrics which can be poem also, corresponding tune is composed , i.e, generating a music for the song, is performed accordingly. The Tune Composition is performed using Neural Networks with Keras library in python. Once the music is generated, a special text-to-speech algorithm is used to sing along the lyrics according to the music .

Contents

1 Introduction **4**

2 Motivation **6**

3 Problem Statement **6**

4 Literature Review **7**
4.1 Microsoft NLP feature detection 7
4.2 Emotelligence on Text, University of Tokushima, IEEE 2007 8
4.3 Detecting Emotion in Text 9

5 Proposed Methodology **11**
5.1 Predicting the emotions based on the inputs: 11
5.2 Generating Suitable Music 13
5.3 Text to Speech Synthesis 15
5.4 Music Superimposition 16

6 Result Analysis **17**
6.1 Phase 1: Emotion Detection from Text 17
6.2 Phase 2: Music Generation 19
6.3 Speech Synthesis and Music Superimpose 21

7 Requirements **21**
7.1 Datasets 21
7.1.1 Dataset for Text-Emotion Classification 21
7.1.2 Dataset for Music Generation 21
7.2 Hardware and Software Requirements 21

8 Future Work and Discussion **22**

9 Conclusion **22**

10 References **24**

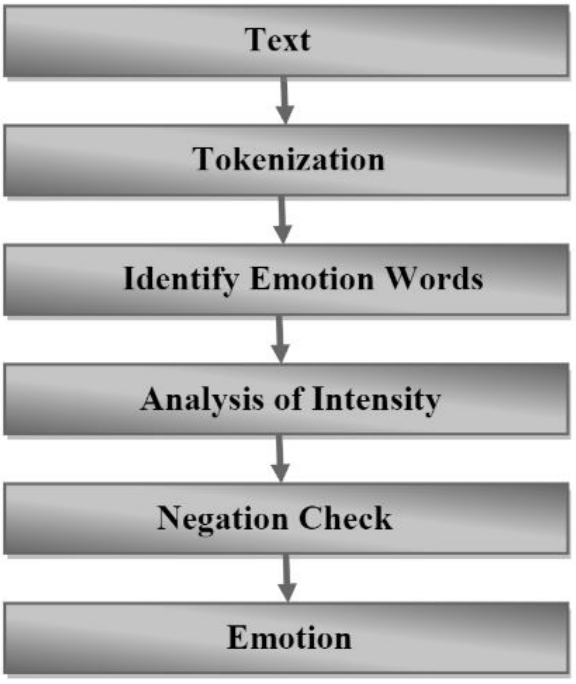
1 Introduction

Music is an art of expressing emotions through frequencies.It is a source of entertainment that combines various sounds together to produce a form that people like. It is much more than just a source of entertainment for most of us. Infact, it helps us express ourselves, calms our soul and boosts our self-confidence. We have seen some amazing music composers who have composed great music with a great deal of underlying musical structure. With the advancement of computer science, now with the help of machine learning algorithms even computers can create such musical structures.

In this paper, we propose a neural networks based system for generating a song based on a given lyrics. The user will give the lyrics of the song to the machine and the machine will be capable of generating music for the song based on the emotion of the lyrics.

The lyrics given by the user can be classified to various moods such as happy, sad, angry, relaxed, etc. We propose to do this using NLP(Natural Language Processing) algorithms. We may use CBOW (Continuous Bag Of Words) technique to identify the emotion based on the words in the text. Based on the detected mood, a neural network should be able to generate a tune with a trained dataset of various types of music. We plan to implement it by using CNN (Convolutional Neural Networks) and LSTM(Long short-term memory) neural networks which are efficient generating models.

The music generated will then be passed to another complex neural network based text-to-speech algorithm which produces a voice to sing verses along in a state of harmony with the music. This algorithm will change the pitch and stress in the voice based on the ongoing music.



Emotion Detection from Text

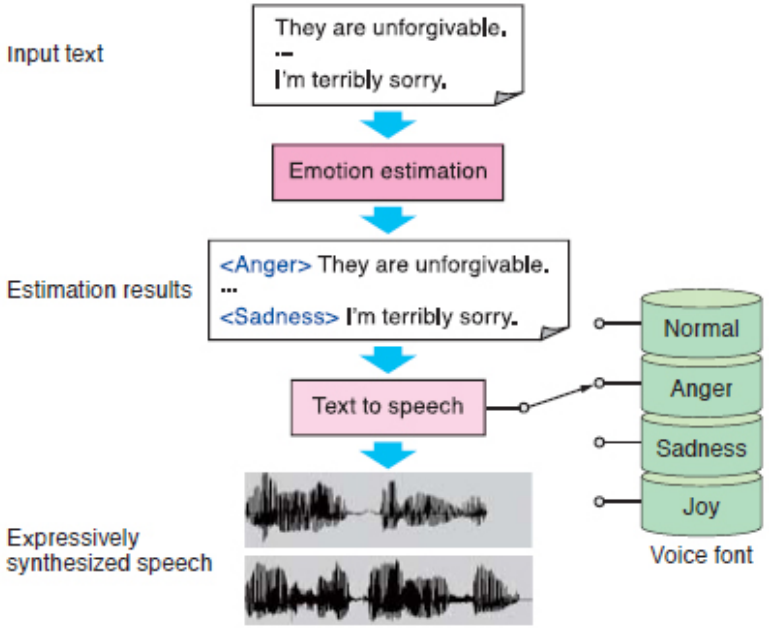


Figure :A general model to generate speech from text

2 Motivation

We aim to explore new possibilities and incorporate them in our everyday lives to make it more comfortable and easy going.Machine Learning had played a great role in the advancement of human lives.The use of machine learning algorithms to predict emotions from a text or to generate some music is something which is very new in the market and there are not much applications which can use this to produce fruitful results.

It consumes a lot of time to understand the lyrics of a song with its emotion and genre and composing music for it.We aim to solve this problem of manually composing music for some given poem or lyrics and then infuse that with the given lyrics to present the required song.

3 Problem Statement

We aim to implement an efficient system for composing music for given poem or lyrics and then infusing it with the text to present the required song . The input will be the poem or the lyrics and the system will generate music for it as well as will try to compose to the complete song with that music and lyrics.

The problem is further divided into 4 sub-problems:

- Predict the emotions/genre from the text.
- Generate music based on the emotions/genre.
- Convert Text to Speech.
- Impose the generated music with the lyrics.

4 Literature Review

4.1 Microsoft NLP feature detection

During the emotion detection process the textual data is classified into emotions like happy, relaxing, angry, sad. This involves Natural Language Processing algorithms. NLP is on the pace now, a lot of research is already being done in this domain, specifically in Sentiment Analysis. Various tools and algorithms are being developed for NLP, making it more flexible for tasks like ours.

1. In [1] the author has used the Microsoft's internal Machine Learning tool to classify a textual data to various emotions. A multi-layered neural network with 3 hidden layers of 125, 25 and 5 neurons respectively, is used to tackle the task of learning to identify emotions from text using a bi-gram as the text feature representation.

2. In [2] the authors have developed an architecture to extract lyrics from an audio file and then classify the song based on the lyrics. They have used 2 types of features: features based on existing frameworks like Jlyrics, Synesketch and ConceptNet (FF) and BOW features. They used a dataset of 903 audio excerpts organized into five clusters, similarly to the MIREX campaign. This dataset and user annotated clusters were gathered from the Allmusic database.

During music generation process the neural network should take an emotion as an input and generate music based on the previously trained music datasets. This involves advanced neural networks like Recurrent Neural Network, Long Short Term Memory neural networks, etc. This field has recently come into picture because till recent times we don't really have the capability or tools to handle music.

1. Recently, Google has launched Magenta, an open source research project exploring the role of machine learning as a tool in the

creative process. Magenta contains various tools which include utilities for manipulating source data (primarily music and images), using this data to train machine learning models, and finally generating new content from these models .

2. In [3], the author has proposed various machine-learned models which converts text to the music. The music generation model first extract one or more structural features from text given as input, then this model generates a musical composition in respond to these structural features of the input text.
3. In [3], the author is using text classification to get the emotions and other sentiments using the DNN(Deep Neural Networks) and then uses it to generate suitable music for the input text.

4.2 Emotelligence on Text, University of Tokushima, IEEE 2007

This paper which was named as the **"Emotelligence on Text"** was published in IEEE 2007 by the Faculty of Engineering of University of Tokushima. This project takes the text as the input from the user and the derive the emotions from it.

This journal has taken 8 basic emotions which are **"joy, trust, surprise, fear, sad, anticipate, disgust and anger"** and treat them as the 8 basic classes for the classification job. These basic emotions can also be used to detect some other complex emotions like love.

The journal has trained their model using the datasets which has collected its information using social networking sites, blogs, etc. First they had tried to extract the important words from it like noun, verb, adjective, etc then they assigned some numerical values to it so that they can train their model on the basis of it.

The results which they got for their model has overall accuracy of the 71 percent. This journal has used 1800 corpus for the purpose of the training and 200 for the testing job. For emotion of joy their accuracy was 96 percent which was very high while for the surprise they achieved only 2 percent accuracy. Their overall result for all the 8 kind of emotions was shown in figure below.

Class	Training set	No. Tested	Correct Output	Wrong Output	Accuracy
Angry	225	25	21	4	84%
Anticipate	225	25	21	4	84%
Disgust	225	25	14	11	56%
Fear	225	25	22	3	88%
Joy	225	25	24	1	96%
Sadness	225	25	20	5	66%
Surprise	225	25	5	20	2%
Trust	225	25	23	2	92%

Figure: Overall result by Emotelligence on Text journal

4.3 Detecting Emotion in Text

This paper named as the **"Detecting Emotion in Text"** was published by Kaitlyn Mulcrone of University of Minnesota, Morris. This journal presented an overview of the various approaches used for detecting emotion from text. It also compared the various approaches and pointed out the similarities and dissimilarities between them. Further it covered the difficulties and challenges faced in the field and also talked about the future of the field in general.

The journal analysed two approaches which included a total of 4 methods. The emotions included in the research were: **anger,fear,joy and sadness**. The result of these tested methods are shown in the following table:

Methods	SemEval			ISEAR			Fairy Tales		
	Prec.	Rec.	F1	Prec.	Rec	F1	Prec.	Rec	F1
MCB	0.077	0.250	0.118	0.100	0.250	0.143	0.102	0.250	0.145
LSA	0.363	0.348	0.340	0.484	0.282	0.228	0.662	0.640	0.630
PLSA	0.189	0.282	0.219	0.260	0.317	0.270	0.282	0.307	0.280
NMF	**0.523**	**0.506**	**0.505**	0.461	0.258	0.166	**0.747**	**0.731**	**0.733**
VAD	0.466	0.422	0.386	**0.528**	**0.417**	**0.372**	0.530	0.404	0.419

Table 1: Overall average results for the three datasets using precision, recall, and f-score; best results are in bold [4].

Here,precision is the ratio of number of correctly labeled sentences retrieved to the total number of sentences retrieved. Recall is the ratio of the number of correctly labeled sentences retrieved to the total number of sentences annotated as correct. After calculating precision and recall,f-score is calculated using the following formula:

f-score = $2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})$

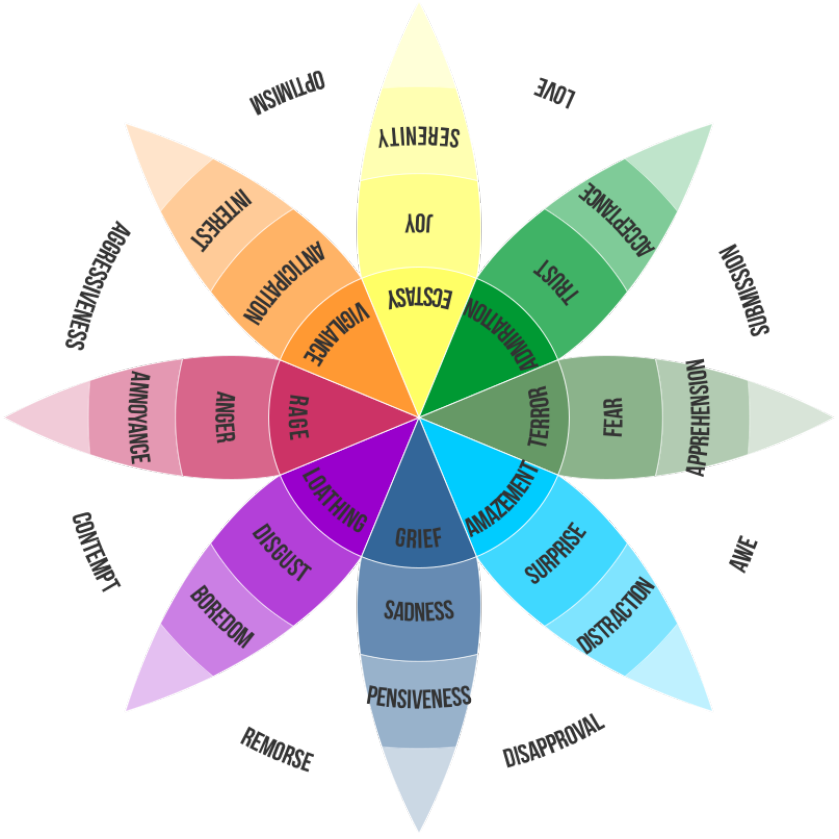
These results show that using VSM with the NMF method produces the best result. It is also seen that the precision,recall and f-score values are higher for fairy tales compared to other datasets.This was because of the way fairy tales dataset was written, the sentences used in it were longer and contained more emotional words than others.

Thus this paper compared different approaches used for emotion detection from text and figured out the best one from them.

5 Proposed Methodology

5.1 Predicting the emotions based on the inputs:

This seems quite similiar to a text classification problems, a contextual based on DNN (Deep Neural Networks), we will be using a simple Tensorflow Hub text embedding module which will train a simple classifier with a given baseline accuracy, we will then analyse the results from our predicted outputs to ensure that the accuracy is achieved at its best. Since DNN's are data hungry, so we will need a large bundle of data sets for accomplishing our task, these datasets will serve as a text classifier to resolve our problem.



Emotions that can be extracted from texts.

TF-Hub provides a feature column that applies a module on the given text feature and passes further the outputs of the module.
Steps done do detect emotions from text are :

1. Collecting a batch of sentences from the lyrics into a tensor of single dimensional strings as input.
2. Preprocessing of the sentences e.g Removal of punctuations and splitting of sentences.
3. Text tokenizer is used to tokenized i.e. top 20000 most frequently used words will be used from the dataset.
4. These words then will be fitted against the input words and every other words with lesser frequency will be removed.
5. A sequence of words is generated .
6. Training set is initialised as these tokens and padded so that each sequence has the same length .
7. As the part of estimator, we will use a CNN classifier.
8. Now , a neural network model is built with 4 layers .
9. First layer is input layer which has

    (a) embedding - changing the discrete category into vectors of continuous numbers .
    it takes max features i.e. top most words as an input vector and gives output of 50 dimensional vector of a fixed maximum size.

    (b) conv1d - it is used to reduce the dimensionality of the vectors and also in such a way that each next layer contains a values depended upon the some of the inputs .

    (c) dense - output layer is used as such to predict the output as in feedforward neural network.

10. The given neural network is trained for epoch value till an accuracy of 90 percent is reached . This trained model is then utilised to predict the emotions from text which is the most probable estimation of the input text based on the features extracted from the input dataset .

5.2 Generating Suitable Music

Generating a suitable music from the known lyrics is kind of a difficult task to work on since we first need to predict the emotions based on the meaning of our lyrics and then generate the music accordingly.
Music is based on sequentially occurring patterns in time. Hence, the network used to achieve this has to be a Recurrent Neural Network. This type of network generates a pattern of sequences, i.e., predicts the next output taking in the input as the previous output. However, using a Recurrent Network does not aid us with full optimality/efficiency. RNN generates the output based only on the previous output. However, music is more complex than just adjacent time steps' dependency. Hence, we use a Long Short Term Memory network. LSTMs store a "history", which is updated and given as the input along with the previous output to the next input, which makes this is highly suitable for music.
Below is the

1. We first give the lyrics of the song on our own , the lyrics provided will be running through the convolution neural networks, the neural network will provide the output for the resulting emotion that can be obtained with the maximum accuracy level , it averages to 3 epochs (iterations) and then decides which emotion best fits for the given lyrics inputted.

2. First, a data set with emotion/genre tagged music will be selected. Then the dataset is extracted as single charecters.

3. Then the data is trained against a neural network model which

has four layers, consisting Embedding Layer, two LSTM layers and an Output Layer.
But the effieciency achieved is not optimal and converges very slowly. Thus an update is required.

4. The weights saved in previous model have been saved and loaded to a second updated model which contains total of five layers, in which an extra LSTM layer is added to the first model. This is done to increase the efficiency of the model.

5. Next, the data will be simplified, extracting only a single track. The track is then converted to an encoding which represents the music on a continuous timescale, together with the genre of the music.

6. This will be accomplished by providing the network with an input sequence, on which the network will make a prediction. The prediction is then appended to the input sequence, and the new input sequence is fed back into the network again for it to make a new prediction. By repeating this process, the network can generate sequences indefinitely.

The LSTM is the main driving force behind the network, it has the ability to learn long-term dependencies embedded in the sequences that flow through it. This is achieved by the extra cell state of the LSTM, it allows previous values to retain in memory.

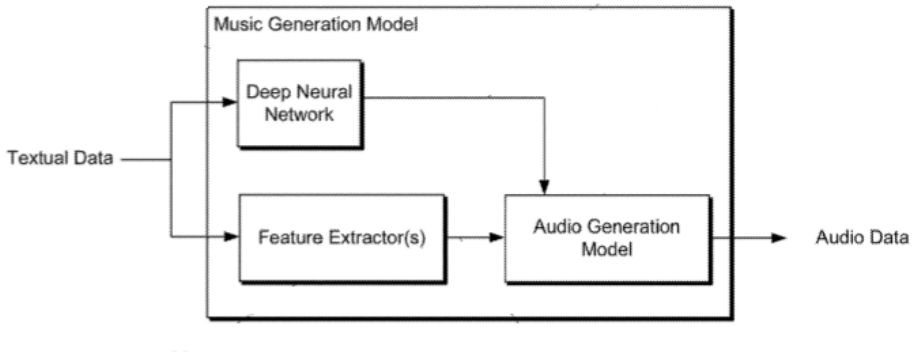


Figure 2 : General model for generating audio for input texts [3].

5.3 Text to Speech Synthesis

The working on text to speech was comparatively easier since we just had to make it feel like someone singing a song . So our approach for achieving it was to make the text or the lyrics inputted to be converted into simple speech using the already predefined gTTs (Google text to speech) library. This library allows us to input any lyrics stream on our own and it just generates the google automated voice based on our input.So finally we are able to perceive one mp3 file based on our input.
A neural network is trained on the data set which is classifying the song to its various emotions using the CNN . The resulting output for the emotion will generate the best fit music from the emotion that is suited for the lyrics. The synchronization was important since it should run for the same duration as that of the maximum time it is taking for running.

5.4 Music Superimposition

Music superimposition is an important phase in the entire project since we need to make the song more realistic as possible. But we are not emphasizing so much on the speech but on the synchronization of the tune generated with the speech so that it will be a more better music generated. For achieving the task ,it was divided into two phases : First phase was to convert the given .mid music file generated to .mp3 music file generated this was achieved using web scrapping tool called selenium ,using scrapping in python , we tend to browse automatically a site onlineconverter.com/midi-to-mp3 using the scrapping tool , our generated midi file is getting automatically uploaded and then converted and downloaded to the corresponding .mp3 format file , So now we have 2 files. The second task is to superimpose the two mp3 files that we have generated using the music ,this was achieved using the pydub and the audio-segment library which is a most common library to superimpose the two music , it generates a wave from the both music files and then superimpose them one over the other, this makes the complete music file.

6 Result Analysis

6.1 Phase 1: Emotion Detection from Text

In the given dataset, we had 14 features which was then merged on their similarity of emotions with others to form a newer dataset of 5 featured emotions on which the model was trained to predict the emotion feature from the input text .

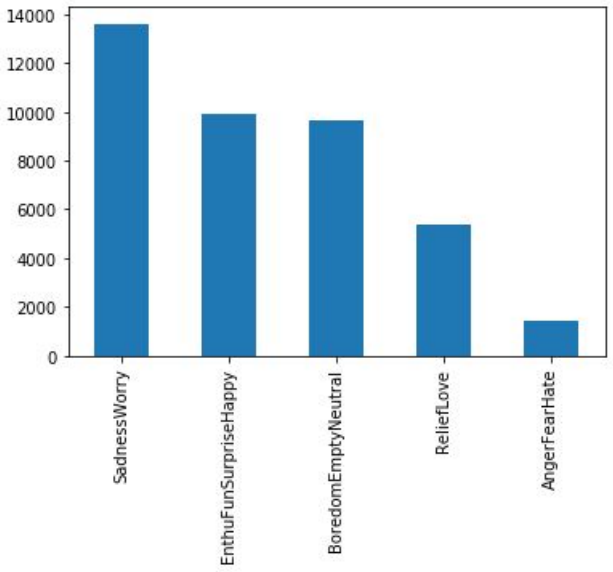


Figure : Input features versus their occurance

After model was trained , we had the accuracy of 95 percent with our model which was later used to predict the text emotions out of it . This is very good in response to what the input has been fed to predict the output of the input text by the model.

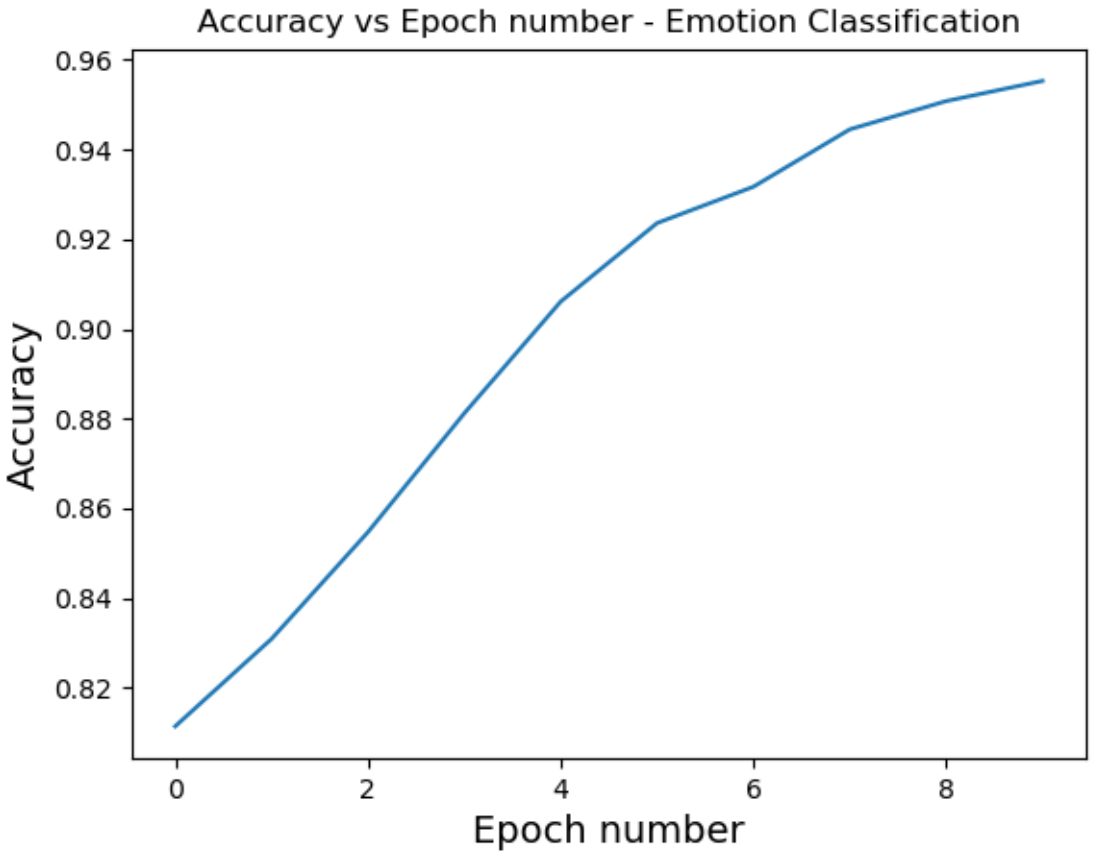


Figure : Epoch Number vs Accuracy

As it is evident from the graph at the end of epoch 10 , the accuracy reached the level of 95 percent which was used later to predict the emotions from text of maximum length 400 which was a constraint.

6.2 Phase 2: Music Generation

In this project , several different music emotion features were trained for the different datasets that were procured manually and curated overtime through various sources .
With the lstm model , output of epoch vs accuracy of some the input music emotions generated are as

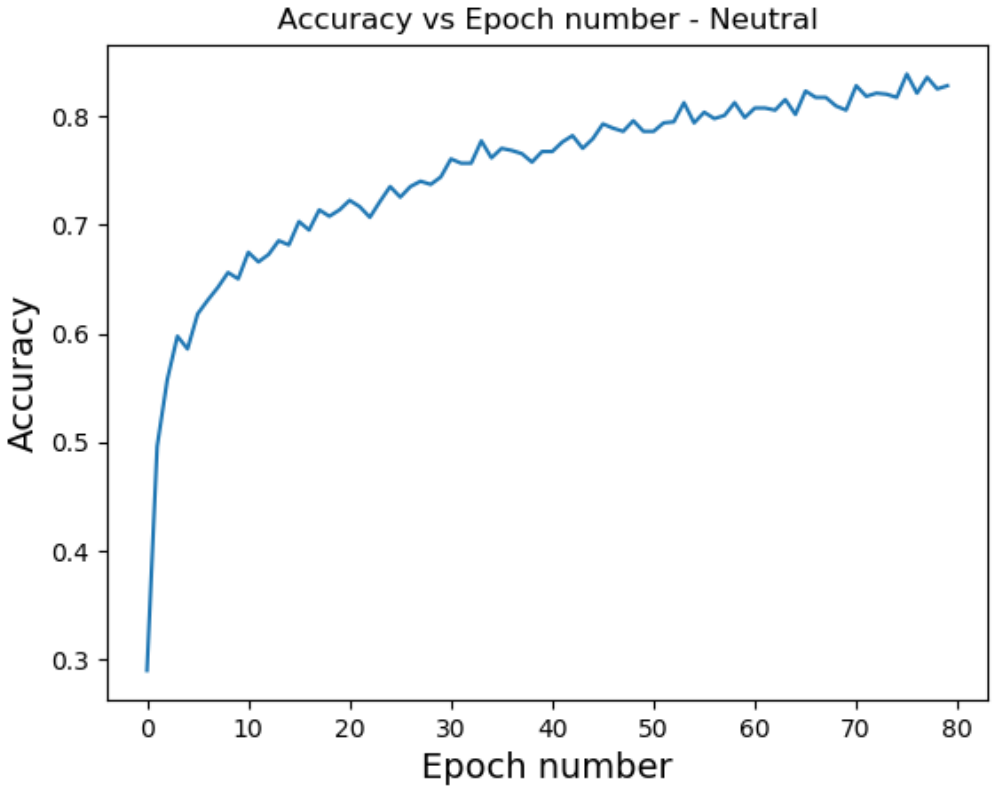


Figure: epoch vs accuracy for the emotion - neutral

At the end of music generation for the neutral dataset , we had close to 85 percent accuracy at the end of 80th epoch .
Below are the result for the model trained for different emotions such as sad , ecstacy and neutral .

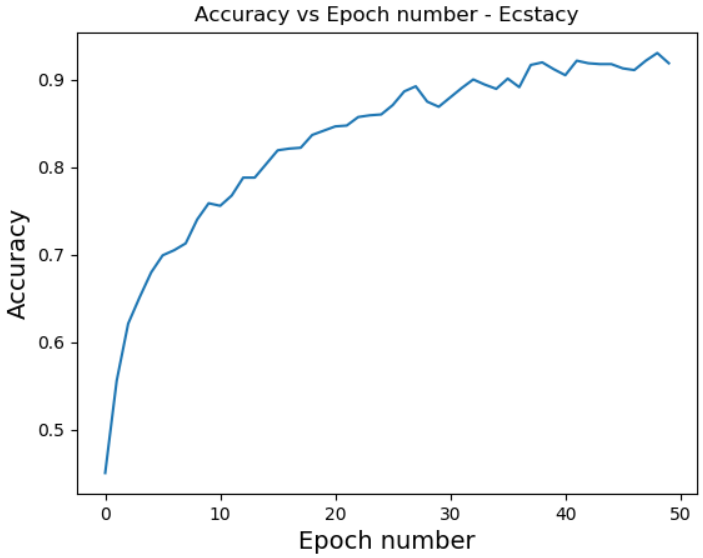


Figure : epoch vs accuracy for the emotion - ecstacy At the end of music generation for the ecstatic dataset , we had close to 90 percent accuracy at the end of 50th epoch .

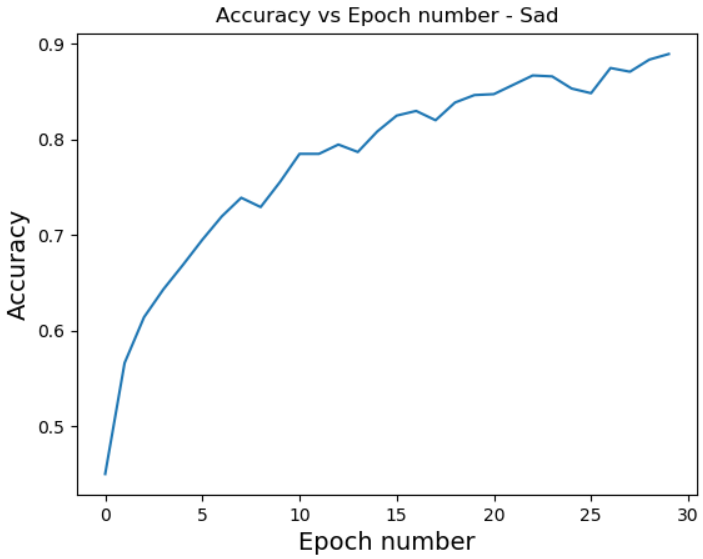


Figure : epoch vs accuracy for the emotion - sad At the end of music generation for the sad dataset , we had close to 90 percent accuracy at the end of 30th epoch.

6.3 Speech Synthesis and Music Superimpose

Midi files are generated from emotion detected from the text input and speech synthesis is done through gtts which is then superimposed to get the output song file .

7 Requirements

7.1 Datasets

7.1.1 Dataset for Text-Emotion Classification

In this part of project , we used publicly available dataset : **text_emotion.csv** from kaggle datasets which has 39k tweets with sentiments labelled in 14 category. This dataset was later merged into 5 emotions namely - disgust , love , neutral , ecstacy and sad .

7.1.2 Dataset for Music Generation

In this part of project, there was no publicly available dataset which can be used for the part of the project , hence data collection and curation needed to be done through variuos online resource and then merging them all to make a dataset of our own with music midi files with labelled emotions along with that .

7.2 Hardware and Software Requirements

- Intel i7 or above processor.
- RAM - 8GB minimum.
- Operating System - Linux (Ubuntu), Windows

- Python 2,3
- Keras and Tensorflow Libraries.

8 Future Work and Discussion

This paper discusses the neural network to generate the music from the input text. This is achieved through the classification of the text into the emotion through the use of the convolutional neural networks and then based on the emotion a tune or music is generated using the LSTM (Long short-term memory). A text to speech conversion has been done to generate a speech for the input text. The speech generated is then superimposed and synchronized with the tune to generate a complete song.

The tune generated is based on the emotions of the input text but the emotion based speech generation has not been achieved fully. There are many possibilities to work on through which emotion based speech and song can be achieved. This can be achieved either by using the datasets with respect to the voice of the particular singer, the way he/she sings, the way he/she gives stress on particular words. The other possibility is that one could try to find out waveform pattern and beats of the tune generated and then try to make the changes in the waveform of the speech accordingly. Like for the fast beats word can be contracted while in case of slow beats stretch on the word could be put.

9 Conclusion

After putting so much effort into our work, building a given sentimental analysis and working upon the final refactoring of the combined code, we finally come to a conclusion of the work , the work mainly

confronts into putting the idea of deep neural networks for building an emotion based detection based on the lyrics inputted , our neural networks easily understands the lyrics code and adjust the weights based on several iterations. Also understanding the song, it will be providing us with the most appropriate tune that can be generated using the work. Some modifications based on synchronizing the song with the speech converted to make it feel more like a music. It was an effective hard work of the entire five members team for the project , and we learnt and explored many new area in the field of machine learning ,like CNN, LSTM and many more.

10 References

1. **Detecting Emotion in Text**
https://github.com/tpsatish95/emotion-detection-from-text/blob/master/literature_survey/Other/Emotions/Detecting%20Emotion%20in%20Text.pdf

2. **Emotion Detection and Recognition from Text Using Deep Learning**
https://www.microsoft.com/developerblog/2015/11/29/emotion-detection-and-recognition-from-text-using-deep-learning/

3. **Ricardo Malheiro,Renato Panda,Paulo Gomes,Rui Pedro Paiva:**
Music Emotion Recognition from Lyrics: A Comparative Study
http://www.ecmlpkdd2013.org/wp-content/uploads/2013/09/MLMUMalheiro.pdf

4. **Dominik Roblek , Douglus Eck : Machine Learning to generate music from text** : United States Patent Application Publication ,Pub No. : US2018/0190249 A1 .

5. **Gabriel Meseguer-Brocal , Alice Cohen-Hadria ,Geoffrey Peeters : DALI: A LARGE DATASET OF SYNCHRONIZED AUDIO, LYRICS AND NOTES**: Proceedings of the 19th ISMIR Conference, Paris, France, September 23-27, 2018.

6. **Music Generation from Deep Learning**
https://medium.com/datadriveninvestor/music-generation-using-deep-learning-85010fb982e2