**Audiobook.AI**

Submitted in partial fulfillment of the requirements

of the degree of

**B. E. Computer Engineering**

By

|  |  |  |
| --- | --- | --- |
| **Daniel Lobo** | **07** | **172028** |
| **Jenny Dcruz** | **16** | **172047** |
| **Smita Deulkar** | **19** | **172113** |
| **Leander Fernandes** | **20** | **172061** |

Guide:

**Ms. Priya Karunakaran**

Assistant Professor



Department of Computer Engineering

St. Francis Institute of Technology

(Engineering College)

University of Mumbai

2020-2021

**CERTIFICATE**

This is to certify that the project entitled **“Audiobook.AI”** is a bonafide work of **“Daniel Lobo” (Roll No. 07), “Jenny Dcruz” (Roll No. 16), “Smita Deulkar” (Roll No. 19), “Leander Fernandes” (Roll No. 20)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.E. in Computer Engineering.

**(Name and sign)**

**Guide**

**(Name and sign) (Name and sign)**

**Head of Department Principal**



**Project Report Approval for B.E.**

This project report entitled ***Audiobook.AI*** by ***Daniel Lobo, Jenny Dcruz, Smita Deulkar, Leander Fernandes*** is approved for the degree of ***B.E. in Computer Engineering.***

Examiners

1.---------------------------------------------

2.---------------------------------------------

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

-----------------------------------------

(Signature)

-----------------------------------------

(Name of student and Roll No.)

Date:

**Abstract**

Music plays a very important role in people’s lives. The music industry is worth $19 billion however the average person doesn’t have a vast knowledge of music theory. This makes it difficult for these people to come up with original and catchy melodies. Automating the creative process of the human means that companies can get multimedia products faster and cheaper. That is why music generators are such an important application of machine learning techniques. With the development of deep learning, neural networks are increasingly used in various art fields such as music, literature and pictures, and the artwork created can be even comparable to humans. Using a deep neural network we can train a model to generate new melodies. Some of the different approaches that can be used to generate music using neural networks are LSTMs, bidirectional RNN, etc. Using Principal Component Analysis we can adjust the component sliders to generate different melodies. This system has multiple applications. For example, musicians can use this application to generate new songs using the unique melodies that our system creates. In audiobooks, using sentiment analysis we can generate emotionally relevant background music. For our project, we will be focussing on creating an application that can take the text of a book as the input and generate relevant background music based on the predicted sentiment.

**Contents**

|  |  |  |  |
| --- | --- | --- | --- |
| **Chapter** | | **Contents** | **Page No.** |
| **1** |  | **INTRODUCTION:** |  |
| **1.1** | **Description** |  |
| **1.2** | **Problem Formulation** |  |
| **1.3** | **Motivation** |  |
| **1.4** | **Proposed Solution** |  |
| **1.5** | **Scope of the project**. |  |
| **2** |  | **REVIEW OF LITERATURE** |  |
| **3** |  | **SYSTEM ANALYSIS** |  |
| **3.1** | **Functional Requirements** |  |
| **3.2** | **Non Functional Requirements** |  |
| **3.3** | **Specific Requirements** |  |
| **3.4** | **Use-Case Diagrams and description** |  |
| **4** |  | **ANALYSIS MODELING** |  |
| **4.1** | **Data Modeling** |  |
| **4.2** | **Activity Diagrams / Class Diagram** |  |
| **4.3** | **Functional Modeling** |  |
| **4.4** | **TimeLine Chart** |  |
| **5** |  | **DESIGN** |  |
| **5.1** | **Architectural Design** |  |
| **5.2** | **User Interface Design** |  |
| **6** |  | **IMPLEMENTATION (if any)** |  |
| **6.1** | **Algorithms / Methods Used** |  |
| **6.2** | **Working of the project** |  |
| **7** |  | **CONCLUSIONS** |  |

Appendix

References

Publications by your group (if any)

Acknowledgements

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Fig. No.** | **Figure Caption** | **Page No.** |
| 1.1.1 | Architectural Diagram |  |
| 1.4.1 | Workflow of Client-end Software |  |
| 1.4.2 | Workflow for training Text-based sentiment analysis model |  |
| 1.4.3 | Workflow for training Audio-based sentiment analysis model |  |
| 1.4.4 | Workflow for Final sentiment analysis model |  |
| 1.4.5 | Workflow for training sentiment relevant Music Generation Model |  |
|  |  |  |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Table Title** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**List of Abbreviations**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Abbreviation** | **Expanded form** |
| i | TBSA | Text-Based Sentiment Analyser |
| ii | ABSA | Audio-Based Sentiment Analyser |
| iii | DFD | Data Flow Diagram |
| iv |  |  |
| v |  |  |

**Chapter 1**

**Introduction**

Oral storytelling is a thousand years old tradition and still abides till this day through digital recordings like podcasts, audiobooks, etc. An audiobook essentially is a recording of a book, novel, story, or other work being read aloud. They have been used extensively to teach critical listening, model good interpretive reading as well as aid students to understand books above their reading level and so on.

* 1. **Description**

Audiobooks are being used on a regular basis by hundreds of users. The system in this report aims to develop emotionally relevant music for preexisting audiobook recordings. The user will enter an audiobook MP3 file as an input to the system. This audiobook will then go through two processes, simultaneously.

* First, the input audiobook will be run through a transcription tool to extract the text from the audiobook. This text will then be analysed using a Text-Based Sentiment Analyzer (TBSA).
* Concomitantly, in the second process, the features of the audio from the audiobook that is given by the user, will be extracted. The audio features are then analysed by an Audio-Based Sentiment Analyzer (ABSA) that will predict the emotions being conveyed in the audio.

Now the system will have obtained 2 values (sentiments) predicted by both, the TBSA as well as the ABSA. The values may vary and lead to an error. To avoid this, the weighted average of values will be calculated in order to generate the final predicted sentiments. Utilizing these predicted sentiments as well as the music generation model that has been explained ahead in this report, our application generates a seamless, distinctive musical score for every segment. These scores are stitched together along with the input audio file to provide the user an audiobook with felicitous background tunes.

Fig1.1.1 Architectural Diagram

* 1. **Problem Formulation**

The speech recordings that exist in the form of audiobooks are relevant but not as impactful. One may seldom find audiobooks with background music. Ordinarily, audiobooks consist of a monotonous speaker that reads the script to a book. Even though the content is being spoken out, the sentiments portrayed are not understood by the user. For example, a sarcastic line, if spoken without voice modulation may be perceived as offensive rather than humorous. Therefore, adding a musical score to the audiobooks will enhance its effectiveness and make a difference in the user’s experience.

Crafting a musical score generator involves smoothly resequencing, looping, and timing the music to match the emotions in the story as they change over the course of the narrative. This is a challenging task even for experts. Due to this most audiobooks today only consist of speech. Existing audio editing tools force story producers to manipulate speech and music tracks using low level waveform editing which are expensive at times.

* 1. **Objective**

To develop a system that generates an apt, emotionally pertinent, unique musical score for an audiobook automatically based on the current narrative for the purpose of ameliorating user-experience while being accurate, cost-efficient, and time saving.

* 1. **Proposed Solution**

**1.4.1 Workflow of the Client-end Software**

The overall workflow of the system is represented in figure 1.4.1, following which each model is explained in detail. The input provided by the user will be in the form of an MP3 file of an audiobook. This audiobook is then analysed by the system and divided into several segments of a fixed size. Following this, the sentiment portrayed in every segment will be predicted by the sentiment analysers. The system’s music generation model then generates music with respect to the sentiments predicted. These musical scores are then overlaid with the audiobook to provide the user an output MP3 file consisting of the audiobook in tandem with a musical background.

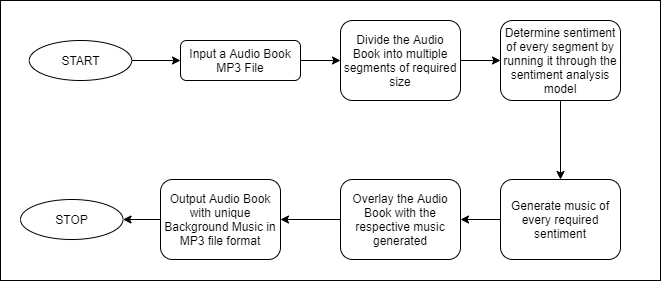


Fig 1.4.1 Workflow of Client-end Software

**1.4.2 Text-Based Sentiment Analysis (TBSA) model**

The TBSA model is one of the two sentiment analysis models of the system. This model runs alongside the ABSA model. The TBSA model is trained using a comprehensive dataset of several texts that consists of relevant labels of different emotions. For instance, a word like ‘Bravo’ will be labeled as happy, ‘unfortunately’ will be labeled as sad, ‘offended’ would be labeled as upset or angry and so on. This model shall be optimized by tweaking the hyperparameters in order to produce the most accurate output. The model with the least loss and highest accuracy will be chosen for the sentiment analysis of the text.

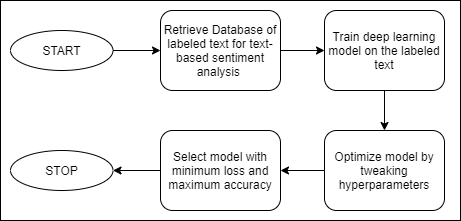


Fig 1.4.2 Workflow for training Text-based sentiment analysis model

**1.4.3 Audio-Based Sentiment Analysis (ABSA) model**

The second sentiment analysis model, the ABSA is trained on a database of MP3 files that are labeled with respect to the emotions that the system takes into consideration. For instance, under the label ‘happy’ an upbeat MP3 audio file will exist. Similarly, for anger one will find the audio file with a high tempo and so forth. The audio features of these files will be extracted so as to train a deep learning model which essentially gives us our ABSA model. This model will be optimized like the previous TBSA model and the one with the least loss and most accuracy will be picked for sentiment analysis.

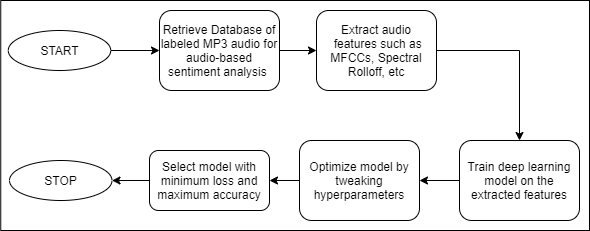


Fig 1.4.3 Workflow for training Audio-based sentiment analysis model

**1.4.4 Final sentiment analysis model**

The final sentiment model fundamentally uses the TBSA and ABSA models. The sentiments predicted by using both these models may or may not be identical. To avoid any oversights, the system takes the weighted average of the predicted sentiments and produces a final sentiment with the most accuracy that is sent ahead to the music generation model.

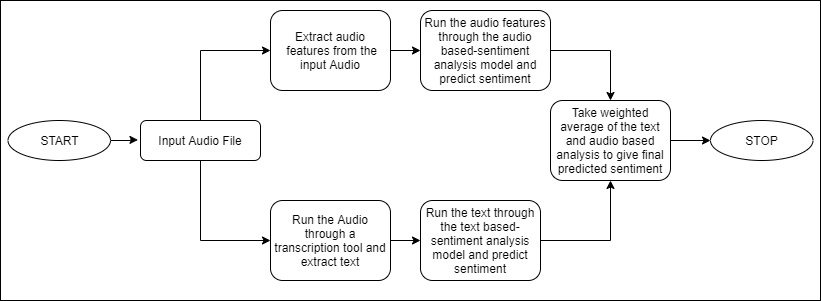


Fig 1.4.4 Workflow for Final sentiment analysis model

**1.4.5 Music generation model**

Besides the sentiment analysis model, the system also consists of a music generation model. In this, we train a GAN model on the extracted audio features of labeled Musical Instrument Digital Interface (MIDI) files that we retrieve from our database. Like we did with the other models, the music generation model is also then optimized by tweaking the hyperparameters. Upon completion of training, we select the model with the slightest loss and most accuracy. Using this model and our final predicted sentiments we generate several musical scores for every segment of the audiobook. These musical scores are stitched together and the music generated is overlaid with the input audiobook file to produce an output audiobook with unique background music that compliments the changes in the plot of the story to create an enhanced immersive experience for the listener.

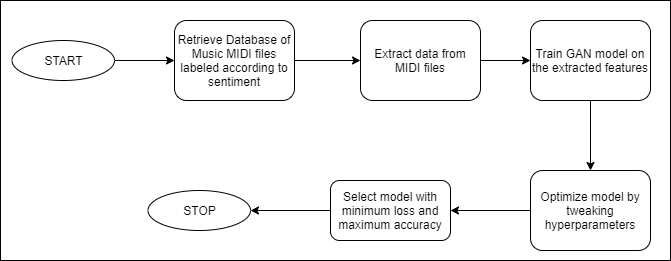


Fig 1.4.5 Workflow for training sentiment relevant Music Generation Model

* 1. **Scope**

This system currently generates music for audiobooks specifically. However, the concept of generating sentimentally relevant music can be applied to other realms as well such as gaming,

**Chapter 2**

**Literature Review**

Review of literature is based on the previous work on Affective Algorithmic Music Composition, more specifically to works that process music in symbolic form in order to generate music with a given emotion. Generating emotionally relevant music for videos and books is a problem that people have attempted to solve in different ways. A common approach for this problem consists of designing a rule-based system to map musical features to a given emotion in a categorical or dimensional space [1]. Steve Rubin and Maneesh Agrawala [2] proposed a system that would resequence music files in order to create a soundtrack that conveys the emotion perceived from an audio book. Their method requires premade music files for different emotions and worked best when the transcript of the audio book was labeled by humans.

**2.1 Sentiment analysis of the audio**

Sentiment analysis is a topic that has been worked on extensively over the past few decades. Most of the work in it focuses on textual data for analysis. This method of analysis may not always accurately recognize the context of the situation described in the text (sarcasm, joy, sadness, etc). Sentiment analysis on audio introduces additional features such as tone, pitch, timbre. But audio sentiment analysis is still in a nascent stage in the research community. Using an hybrid approach comprising of audio analysis as well as text analysis would greatly improve the accuracy with which the system can extract emotions.

Navas, Eva, Inma Hernez, and Iker Luengo [4] have shown that these prosodic features and acoustic features such as power and pitch contribute to the sentiment variation. By incorporating these audio features of the accompanying speech, the capability of the model to recognize the aforementioned context can be increased.

Arpit Shah and Shivani Firodiya [3] used acoustic features such as MFCC, STFT, Contrast, Mel Spectrum, Chroma and Tonnetz extracted from the audio clips of the RAVDESS dataset to train their fully connected DNN model for sentiment analysis on audio.

Classification of sentiments or emotions can be achieved through different measures. A popular approach is Russell's circumplex model [5] which classifies emotions based on their valence and arousal values. The valence dimension indicates whether an emotion is positive or negative, whereas arousal indicates the intensity of the emotion.

**1.2 Generation of Music with sentiments**

There have been multiple studies in the domain of music generation. However, generating music that conveys a certain sentiment has been a major challenge. The goal of Affective Music Composition (AMC) is to automatically generate music that is perceived to have specific emotions. Sentiment is perceived in music due to several features such as melody, harmony, tempo, timbre, etc. Quieter volume, slower tempo and lower pitch is found to convey the emotion of sadness whereas louder volume, faster tempo and higher pitch is perceived as happier. Ferreira and Whitehead [1] have presented a generative mLSTM model that can be controlled to generate symbolic music with a given sentiment. Their model had been trained on a corpus of MIDI files. (MIDI files aren't like regular audio files. They are smaller in size and don’t contain actual audio data. These files explain what notes are being played as they are played, along with the duration and intensity of each note.)

They treated the music composition problem as a language modeling problem and hence represented music pieces as a sequence of words and punctuation marks from a vocabulary that represents events retrieved from the MIDI file. Their model used a Genetic Algorithm (GA) to optimize the weights of their neurons in order to lead the mLSTM to generate only positive or negative pieces. Two independent executions of this GA was performed, one to optimize the mLSTM for generating positive pieces of music and another was for negative pieces. This model could also be used to perform sentiment analysis of symbolic music.

**Chapter 3**

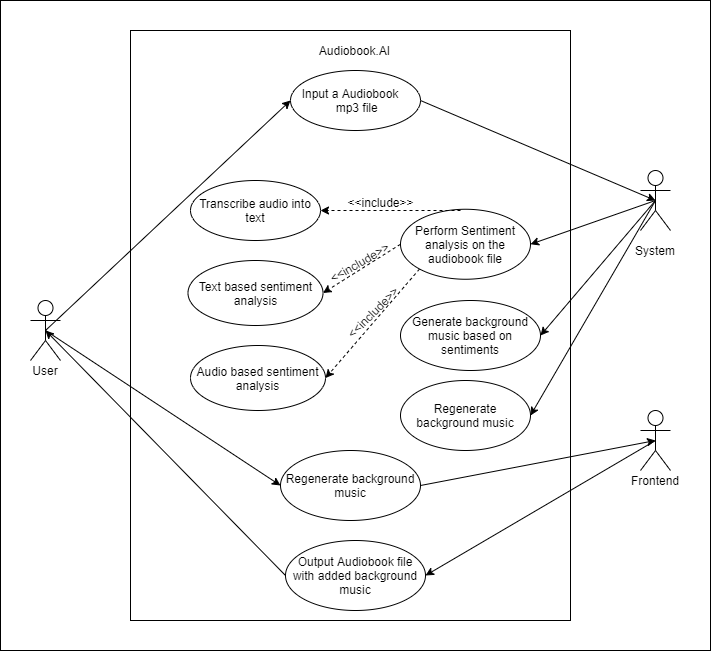
**System Analysis**

**3.1 Functional requirements**

**3.2 Non-Functional requirements**

**3.3 Specific requirements**

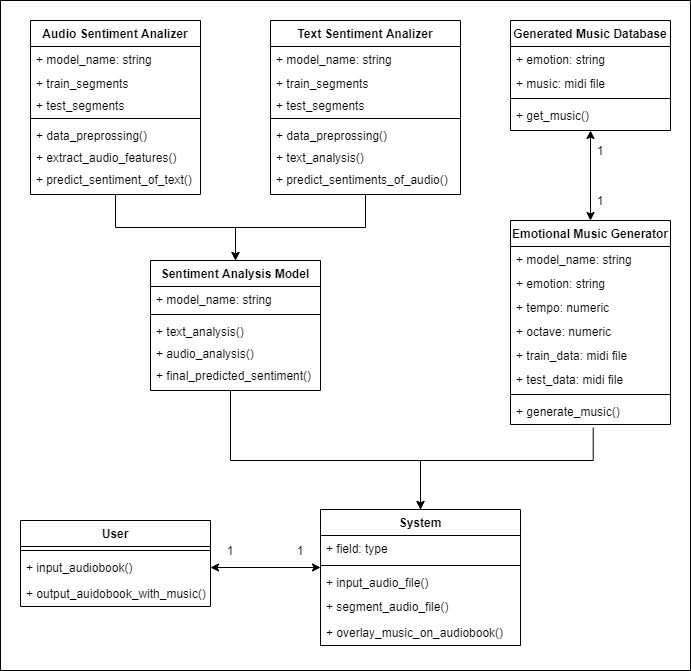
**3.4 Use-Case diagram and description**

****

**Chapter 4**

**Analysis Modeling**

**4.1 Class Diagram**

****

**4.2 Functional Modeling**

Figure 4.2.1 displays the level 0 Data Flow Diagram (DFD) of the Audiobook.AI system. The User gives a MP3 Audiobook file to the system as an input. The sentiments of the audiobook are then analysed. Using these sentiments, emotionally pertinent music is generated and overlaid with the audiobook.

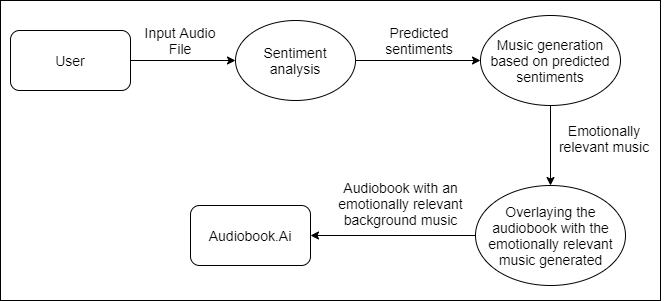
****

Figure 4.2.1 Level 0 DFD

Figure 4.2.2 shows the level 1 DFD of the Audiobook.AI system. It describes the system more elaborately. The input audiobook file goes through 2 processes. In one process the text is extracted from the audiobook, analysed and sentiments are predicted by the TBSA. Simultaneously in another process, the input audiobook file’s features are extracted and the sentiments are analysed by the ABSA. The average of these predicted values are considered for music generation. The music generated is then overlaid with the audiobook and presented as an output to the user.

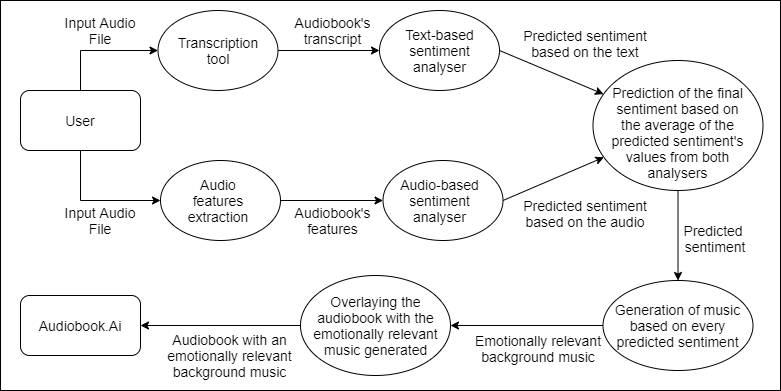


Figure 4.2.2 Level 1 DFD

**4.3 Sequence Diagram**

Figure 4.3 represents the sequence diagram of the Audiobook.AI application. The user inputs the audiobook MP3 file and this information is forwarded to the system which is used to transcribe audio as well as extract features from the audio. The transcribed text and extracted features are then passed through the sentiment analyzer to generate predicted sentiments. The music generator generates emotionally relevant music based on the predicted sentiments. The system overlays the audiobook with generated emotionally relevant music. The system then produces an audiobook with an emotionally relevant background musical score as output to the user.

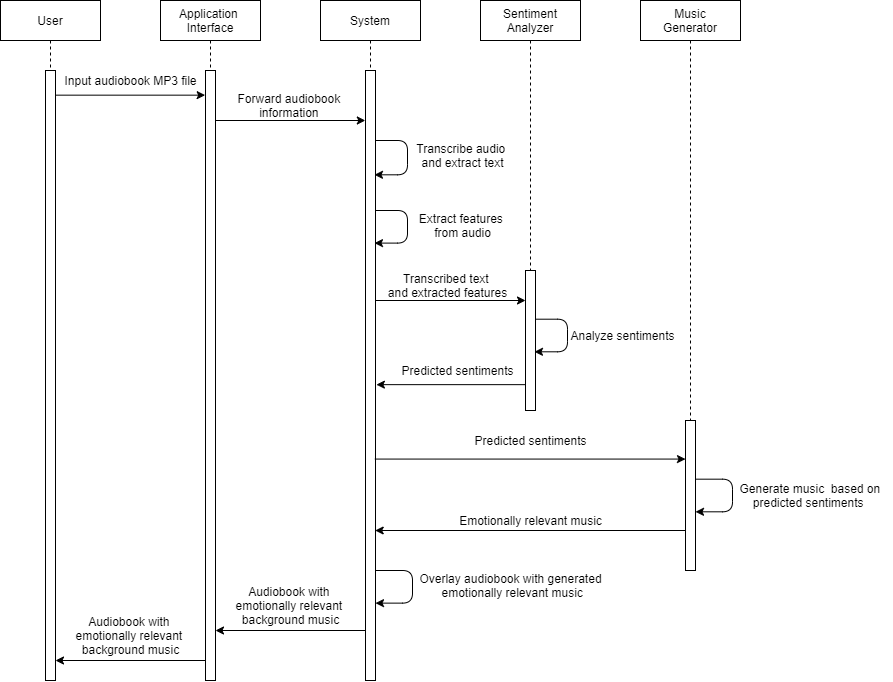
****

Figure 4.3 Sequence Diagram

**4.4 Timeline Chart**

**Chapter 5**

**Design**

**5.1 Architectural Design**

**5.2 User Interface Design**

**Chapter 6**

**Implementation (if any)**

**Chapter 7**

**Conclusion**

**Appendix**

**References**

.

[1] Lucas N. Ferreira, Jim Whitehead. “Learning to Generate Music With Sentiment”, *20th International Society for Music Information Retrieval Conference, Delft, The Netherlands,* 2019.

[2] Steve Rubin and Maneesh Agrawala. 2014. “Generating emotionally relevant musical scores for audio stories”. In Proceedings of the 27th *annual ACM symposium on User interface software and technology (UIST '14).* Association for Computing Machinery, New York, NY, USA, 439–448.

[3] Arpit Shah, Shivani Firodiya, “Audio Sentiment Analysis after a Single-Channel Multiple Source Separation.” Indiana University Bloomington.

[4] Navas, Eva, Inma Hernez, and Iker Luengo. “An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS.” IEEE transactions on audio, speech, and language processing 14.4 (2006).

[5] Russell, J. A. A circumplex model of affect. Journal of personality and social psychology 39, 6 (1980), 1161.

**Acknowledgements**

A project is an activity that needs involvement of a lot of people from different cadres and calibres. Hence it would be highly unfair on our part to claim credit for the whole project. With due respect, we express our sincere gratitude to our esteemed Principal Dr. Sincy George for this teaming experience and for giving us an opportunity to work on this project.

We sincerely want to thank our project guide Ms. Priya Karunakaran for her valuable guidance and advice she has given us when faced difficulties while working on the project. We are also grateful for her constructive criticism and for helping us develop this idea to pursue the project. We would also like to thank our Class Coordinator Ms. Safa Hamdare for her notable inputs and the advice she has given us in our projects.

We would also like to thank our Head of the Department, Dr. Kavita Sonawane and the Computer Department staff who have helped us throughout this project. We are fortunate enough to get encouragement and support from all teaching staff of the computer engineering department. We would also like to extend our sincere regards to all the non-teaching staff of the computer department for their timely support.