Dissertation on

"FeelSpeak: Generating Emotional Speech with Deep Learning"

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE20CS390A – Capstone Project Phase - 1

Submitted by:

M H SOHAN	**PES1UG20CS235**
RAHUL ROSHAN G	**PES1UG20CS320**
ROHIT ROSHAN	**PES1UG20CS355**
S M SUTHARSAN RAJ	**PES1UG20CS362**

Under the guidance of

Prof. V R BADRI PRASAD
Associate Professor

January - May 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100 Feet Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

'FeelSpeak: Generating Emotional Speech with Deep Learning'

is a bonafide work carried out by

M H SOHAN	**PES1UG20CS235**
RAHUL ROSHAN G	**PES1UG20CS320**
ROHIT ROSHAN	**PES1UG20CS355**
S M SUTHARSAN RAJ	**PES1UG20CS362**

In partial fulfilment for the completion of sixth semester Capstone Project Phase - 1 (UE20CS390A) in the Program of Study - **Bachelor of Technology in Computer Science and Engineering** under rules and regulations of PES University, Bengaluru during the period Jan. 2023 – May. 2023. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6th semester academic requirements in respect of project work.

Signature	Signature	Signature
Prof. V R BADRI PRASAD	Dr. Shylaja S S	Dr. B K Keshavan
Associate Professor	Chairperson	Dean of Faculty

External Viva

Name of the Examiners	**Signature with Date**
1. _____	_____
2. _____	_____

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **"FeelSpeak: Generating Emotional Speech with Deep Learning"** has been carried out by us under the guidance of Prof. V R Badri Prasad, Associate Professor, and submitted in partial fulfilment of the completion of sixth semester of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2023. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES1UG20CS235	**M H SOHAN**	_____
PES1UG20CS320	**RAHUL ROSHAN G**	_____
PES1UG20CS355	**ROHIT ROSHAN**	_____
PES1UG20CS362	**S M SUTHARSAN RAJ**	_____

ACKNOWLEDGEMENT

We would like to express our gratitude to Prof. V R Badri Prasad, Department of Computer Science and Engineering, PES University, for her/his continuous guidance, assistance, and encouragement throughout the development of this UE20CS390A - Capstone Project Phase – 1.

We are grateful to the project coordinator, Dr. Priyanka H., all the panel members & the supporting staff for organizing, managing, and helping the entire process.

We take this opportunity to thank Dr. Shylaja S S, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support we have received from her.

We are grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, Dr. B.K. Keshavan, Dean of Faculty, PES University for providing us various opportunities and enlightenment during every step of the way.

Finally, this project could not have been completed without the continual support and encouragement we have received from our family and friends.

ABSTRACT

The goal of this project is to convert normal text to speech with emotion. The project involves two phases: training and testing. In the training phase, we use a text dataset with labelled emotions to train a model that can detect emotions from text. We also use a speech dataset with labelled emotions to train a regression model that can learn the pitch, intensity, and modulation of each emotion. The trained models are then used in the testing phase to generate emotional speech from text.

During the testing phase, the input text is first converted to neutral speech using text-to-speech (TTS) software. The input text is also given to the emotion detection model, which outputs the corresponding emotion. The emotion and the neutral speech are then fed into the regression model, which applies the learned pitch, intensity, and modulation to the neutral speech to generate speech with the desired emotion. The resulting speech with emotion can be outputted as an audio file.

Overall, this project demonstrates how machine learning techniques can be used to create speech with emotion from normal text. By training a model to detect emotions from text and a regression model to learn the speech features of different emotions, we can generate speech that accurately reflects the emotional content of the input text.

TABLE OF CONTENT

LIST OF DIAGRAMS

1. INTRODUCTION

This project aims to convert a given text into speech with appropriate emotional cues. The project consists of two main phases i.e., training and testing.

In the training phase, we start by splitting a text dataset with labelled emotions into 80-20 train-test sets. The training set is used to train an emotion detection model, which uses emotion lexicons to identify which words convey emotions and RNN/RoBERTa/LSTM models to learn the relationship between text and emotions. The test set is used to evaluate the performance of the model.

In parallel, we use a speech dataset with labelled emotions to train a regression model to learn the speech features (such as pitch, intensity, and modulation) associated with different emotions. The trained regression model is then used to apply the appropriate speech features to neutral speech during the testing phase.

During the testing phase, the input text is first converted to neutral speech using a TTS software. The text is then passed to the trained emotion detection model, which identifies the corresponding emotion. The emotion and neutral speech are then passed to the trained regression model, which applies the appropriate speech features to produce the final emotional speech output.

Overall, this project aims to create a more human-like text-to-speech system that can convey appropriate emotional cues in speech.

2. PROBLEM DEFINITION

The problem addressed by this project is the lack of emotional cues in traditional text-to-speech systems. Although these systems can convert text to speech, they often produce monotone or robotic-sounding speech that lacks the appropriate emotional cues to convey the intended meaning of the text. This can lead to misinterpretation and misunderstandings in communication, particularly in situations where emotions play a critical role, such as in customer service, counseling, or entertainment.

The goal of our project is to develop a text-to-speech system that can generate speech with appropriate emotional cues to improve communication and enhance user experience. By incorporating emotional cues into speech, the system can better convey the intended meaning of the text and enhance the overall effectiveness of the communication. This can have significant practical applications in various fields, such as customer service, healthcare, and education.

3. LITERATURE REVIEW

3.1 Literature Review - 1

Title	**Emotionally charged text classification with deep learning and sentiment semantic**
Author	**Dilip K. Prasad; Jeow Li Huan; Chai Quek; Arif Ahmed Sekh**
Year	**2021**
Method	**The methodology involves converting textual documents into a series of vector representations that encode the semantic content of the text. Subsequently, a recurrent neural network is employed to analyze these vector sequences and detect patterns in the long-range relationships between them. To enhance the classification accuracy, sentiment vectors can be appended as a fully connected layer to the word vectors.**
Abstract	**Text classification is a popular tool in natural language processing work, and the most advanced classifiers use the vector space model to extract any required useful features. However, to achieve even greater accuracy, it's important to investigate more complex document representations, such as vector sequences or matrices, that also take into account the emotions expressed in the text.**

Advantages	**The proposed method has achieved a performance that is currently the best in the field. By incorporating sentiment semantics (Using SentiWordNet lexicon) , the accuracy of emotion classification has been improved.**
Limitations	**The proposed method only focuses on emotionally charged text classification, and it may not generalize well to other types of text classification tasks.** **Also the model doesn't deal with pictorial data (like emojis,etc).** **The method requires a significantly huge amount of labeled data for training purposes, which may be difficult to obtain in some domains.**
Sentiment Information	**SentiWordNet (a lexicon) is an extension of WordNet that categorizes words into groups that have synonyms and by giving each group a score that tells us how positive, negative, or neutral the words in that group are.**
Summary	**The following paper's proposed method that utilizes deep learning and sentiment semantics involves representing documents as a sequence of vectors that carry semantic meaning. These vectors are then classified using a recurrent neural network that recognizes long-range relationships. To further improve the classification accuracy, additional sentiment vectors can be easily attached as a fully connected layer to the word vectors. This approach has shown to achieve higher accuracy compared to classical techniques on certain datasets.**
Source	**Refer Bibliography**

3.2 Literature Review - 2

Title	**Emotion Controllable Speech Synthesis Using Emotion-Unlabeled Dataset with the Assistance of Cross-Domain Speech Emotion Recognition**
Author	**Dongyang Dai, Xiong Cai , Zhiyong Wu, Xiang Li, Jingbei Li, Helen Meng**
Year	**2021**
Method	**The proposed method involves training a generative model on an emotion-unlabeled dataset, and then using cross-domain speech emotion recognition to control the emotion of the generated speech.**
Abstract	**The paper presents a method for synthesizing speech with emotions using an emotion-unlabeled dataset and cross-domain speech emotion recognition. The proposed method shows good results in producing speech with desired emotions.**
Advantages	**The proposed method does not require emotion-labeled speech data in large quantities.**

Limitations	**The effectiveness of the proposed method heavily relies on the accuracy of the cross-domain speech emotion recognition. The method may also struggle to generate speech with complex emotions.**
Summary	**The proposed system was evaluated on an emotion-unlabeled dataset and achieved an average Mean Opinion Score (MOS) of 3.78 out of 5 for speech quality and an accuracy of 78.95% for speech emotion recognition. These results demonstrate the effectiveness of the proposed system in generating speech with a wide range of emotions and accurately recognizing the emotion in the generated speech.**
Source	**Refer Bibliography**

3.3 Literature Review - 3

Title	**Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks**
Author	**Teddy Surya Gunawan, Hasmah Mansor, Syed Asif Ahmad Qadri, Taiba Majid Wani, Mira Kartiwi and Nanang Ismail**
Year	**2020**
Method	**The paper suggests two deep learning techniques for emotion recognition from speech. One method involves using CNNs on spectrograms, while the other involves using a DS CNN. The DSCNN processes multiple frames of the spectrogram simultaneously to utilize the temporal structure of the speech signal. Preprocessing of the speech signals is required to obtain spectrograms, followed by training deep learning models on the spectrograms. Finally, the models are utilized to predict the emotion labels of new speech signals.**
Abstract	**The study puts forth two different deep learning approaches for Speech Emotion Recognition - Convolutional Neural Networks (CNNs) and DS CNN. The study then measures their effectiveness on a dataset of speech recordings labeled with emotions. The two approaches involve processing speech signals to obtain spectrograms, training deep learning models on spectrograms, and predicting emotion labels for new speech signals. The study provides an analysis of the strengths and weaknesses of the proposed approaches and compares**

	their performance with other contemporary techniques for SER.
Advantages	**The methodologies suggested by the authors produce outstanding outcomes on a standard dataset for Speech Emotion Recognition. The authors also present an innovative modification of DSCNNs, which leverages the speech signal's temporal structure and can be utilized for other speech-related tasks. The authors conduct numerous experiments to test various configurations and architectures of the deep learning models, providing valuable insights into their respective capabilities and drawbacks.**
Limitations	**Since only one benchmark dataset was used for the evaluation, it's possible that the results won't transfer to other datasets or practical applications.** **The features that the deep learning models have learnt, which might shed light on the Speech Emotion Recognition's underlying workings, are not thoroughly examined in the research.** **The proposed approaches' computational effectiveness is not examined in the research, which might be problematic for real-time applications.**
Feature Extraction	**The network is trained on a dataset of speech recordings with annotated emotions, and is able to learn the patterns of pitch and intensity that are associated with different emotions. Once the network is trained, it can be used to extract these features from new speech signals and classify them into different emotion categories.**

	Specifically, DSCNN uses a series of convolutional layers to extract low-level features from the raw speech signal, such as pitch and intensity. It then uses a series of recurrent layers to capture temporal dependencies between these features over time. Finally, it uses a fully connected layer to map the learned features to the output emotion categories. **By using DSCNN to extract speech features, the authors are able to achieve high accuracy in classifying speech into different emotion categories, which can be useful for applications such as speech synthesis or emotion recognition in human-computer interaction.**
Summary	**Using Deep Stride Convolutional Neural Networks (DSCNNs) , this research presents two strategies for Speech Emotion Recognition (SER) using deep learning techniques. On a benchmark dataset, both approaches achieve cutting-edge performance, and the DSCNN makes use of the temporal structure of the speech signal. Although the evaluation is only done on a single dataset, the paper offers insights into the methodologies' advantages and disadvantages.**
Source	**Refer Bibliography**

3.4 Literature Review - 4

Title	**Contextual Emotion Detection in Text using Deep Learning and Big Data**
Author	**Papel Chandra, Mostasim Billah, Priyadharshini Balaji, Md. Tabil Ahammed, Sudipto Ghosh, Rabiul Hasan Emon, Md Irfan Ahamad**
Year	**2022**
Method	**Input and Preprocessing, Spell Correction , Word Embedding and Model Training.** • **Input data: The first step is to input the dataset for the project.** • **Pre-processing: This step involves cleaning and preparing the data for analysis, such as handling missing or invalid data.** • **Removing invalid words: In this step, invalid or irrelevant words are removed from the text data.** • **Removing extra spaces: Any unnecessary spaces in the text data are removed.** • **Annotated corpus: The text data is annotated to add value to the corpus for future study and development.** • **Segmentation: The text data is divided into segments or sectors based on their significance, openness, notability, productivity, and growth potential. The segmentation can be done based on various criteria, including words, lines, and emoticons.**

	• **LSTM: The preprocessed test is sent to the model to detect emotion from text.**
Abstract	**The paper proposes a model for detecting emotions in text using big data and deep learning algorithms, specifically the LSTM model. The proposed model involves pre-processing the input data by removing invalid words and extra spaces, and segmenting the data based on various criteria. The paper achieved an accuracy of 0.85 in detecting emotions such as happy, sad, angry, and others. However, it is important to carefully consider the pre-processing step as removing too much information can negatively impact the accuracy of the model.**
Advantages	**Removing invalid words and extra spaces can increase the accuracy of the model. The data becomes more standardized and easier to process, leading to better accuracy in the model.** **Every step in the proposed model is independent of each other so we can use any other algorithm other than LSTM to increase the accuracy**
Limitations	**It is important to note that this preprocessing step should be done carefully, as removing too many words or spaces can result in loss of important information and can negatively impact the accuracy of the model.** **The proposed model does not deal with emojis and the emotion the paper detected is very less (happy, sad, angry and others).**

Result	**The paper tries to increase accuracy by carefully preprocess it and removing unwanted inputs to the model.** **The paper reaches a high accuracy of 0.85 accuracy.**
Source	**Refer Bibliography**

4. DATA

Datasets are used in Machine Learning (ML) projects to train, test, and evaluate machine learning models. A dataset is a collection of data points or examples that are used to teach a model to recognize patterns or relationships between the input features and the output labels. The larger and more diverse the dataset is, the better the machine learning model will be able to learn new, unseen data.

Here for our project we have found four dataset where two are for training model 1 where it is used to train the model to detect emotion from text. There is a novel dataset where it is used for testing phase and there is speech dataset for machine learning model 2. We will be describing the dataset below:

4.1. Twitter dataset for emotion detection from text

This dataset has 40,000 rows. In this case, the dataset likely contains tweets that were labeled with different sentiment categories, possibly as part of a supervised learning task in machine learning.

The dataset includes three columns: "tweet_id", "sentiment", and "content". The first column represents a unique identifier for each tweet, while the second column indicates the sentiment expressed in the tweet (e.g., "empty" or "sadness"). The third column contains the actual content of the tweet.

- tweet_id: This column contains a unique identifier for each tweet in the dataset.
- sentiment: This column contains the sentiment associated with each tweet. In this case, it looks like the sentiment can be one of several values, including "empty," which suggests a neutral sentiment or an absence of any strong emotion. It's possible that there are other sentiment categories in the dataset as well, but without more information it's difficult to say for sure.
- content: This column contains the text content of each tweet. In this case, the tweets appear to be personal statements or opinions about various topics, and they may include references to other people or events.

This dataset could be used to train or test a machine learning model for emotion recognition in text. A machine might learn to recognise patterns and forecast the sentiment of fresh text inputs by analysing the content of each tweet and comparing it to the associated sentiment label.

4.2. SemEval2018 - task E-c

The dataset consists of three text files - development, testing and training data set. Each file contains several rows, and each row represents a tweet with its ID, content, and various emotion labels such as pessimism, sadness, anticipation, disgust, anger, love, optimism, surprise, trust, fear, and joy. The emotion labels represent the intensity of the corresponding emotion in the tweet, with 1 indicating the highest intensity and 0 indicating the lowest.

- ID: This column contains a unique identifier for each tweet in the dataset.
- Tweet: This column contains the text content of each tweet, which includes a mention of two Twitter users, and a reference to a past incident involving them.
- pessimism, sadness, anticipation, disgust, anger, love, optimism, surprise, trust, fear, and joy: These columns contain binary values (0 or 1) that indicate whether the corresponding emotion is present in the tweet. For example, the "Anger" column has a value of 1 for this tweet, indicating that the text contains some degree of anger.

By analyzing the content of each tweet and comparing it to the associated emotion labels, a model could learn to recognize patterns and predict the presence or absence of certain emotions in new text inputs. In this case, the tweet expresses some degree of anger and disgust, and does not express any of the other emotions in the dataset.

4.3 Lovely Ilonka novel dataset

The story "Lovely Ilonka" could be used as a small dataset for detecting emotion from text. The dataset consists of a single story with a length of around 600 words and it has 2,46,989 lines with 11.9mb as file size. It includes characters, events, and dialogues that can be labeled with different emotions.

For example, emotions such as disappointment, hope, determination, and love could be associated with the prince's character as he goes on a journey to find the three bulrushes and his promised bride. The old woman's and old man's responses to the prince's questions about the bulrushes could be labeled with emotions such as curiosity, uncertainty, and helplessness. Ilonka's emotions could be labeled with joy when she is freed from the bulrush, fear and desperation when she is thrown into the well, and relief and happiness when she is rescued by the prince.

The dataset could be manually labeled by a person or by using automated techniques such as sentiment analysis or emotion detection algorithms.

4.4. Speech dataset

- AESDD - around 500 utterances by a diverse group of actors (over 5 actors) simulating various emotions.
- ANAD - 1384 recording by multiple speakers; 3 emotions: angry, happy, surprised.
- DEMoS - 9365 emotional and 332 neutral samples produced by 68 native speakers (23 females, 45 males); 7/6 emotions: anger, sadness, happiness, fear, surprise, disgust, and the secondary emotion guilt.
- DES - 4 speakers (2 males and 2 females); 5 emotions: neutral, surprise, happiness, sadness and anger.
- EEKK - 26 text passages read by 10 speakers; 4 main emotions: joy, sadness, anger and neutral.
- Emo-DB - 800 recordings spoken by 10 actors (5 males and 5 females); 7 emotions: anger, neutral, fear, boredom, happiness, sadness, disgust.
- IEMOCAP - 12 hours of audiovisual data by 10 actors; 5 emotions: happiness, anger, sadness, frustration and neutral.
- OGVC - 9114 spontaneous utterances and 2656 acted utterances by 4 professional actors (two male and two female); 9 emotional states: fear, surprise, sadness, disgust, anger, anticipation, joy, acceptance and the neutral state.
- The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) - The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and the song contains calm, happy, sad, angry, and fearful emotions.
- SAVEE Dataset - 4 male actors in 7 different emotions, 480 British English utterances in total.
- TESS - 2800 recordings by 2 actresses; 7 emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.

5. SYSTEM REQUIREMENTS SPECIFICATION

5.1. Introduction

The purpose of this document is to define the requirements for the development of a system that can generate emotional speech from text. The system will utilize deep learning models for text preprocessing, emotion classification, and prosody modeling. The system will take a user-inputted text and output speech with appropriate emotional expression. The system will be designed to be integrated into existing virtual assistant applications or other speech-based applications that require emotional expression in speech.

5.1.1. Project Scope

The purpose of this project is to develop a system that can generate emotional speech from given text inputs. The system will use deep learning techniques to understand the context and emotions conveyed in the text, and then generate corresponding emotional speech in a natural-sounding way.

The benefits of this system include providing a new tool for individuals who have difficulty expressing their emotions verbally, as well as enhancing the emotional expressiveness of various applications such as virtual assistants, chatbots, and video games.

The objectives of the project are to develop a machine learning model that can accurately identify and classify the emotions conveyed in text, to develop a text-to-speech synthesis system that can generate emotional speech with natural-sounding intonation, and to integrate these two components into a working prototype.

The scope of the project is limited to the English language and a set of predetermined emotions such as happiness, sadness, anger, and neutral. The system will be designed to work on a desktop or laptop computer with a standard microphone and speakers, and will not require any specialized hardware or software. The accuracy of emotion classification and the quality of generated speech will be the primary focus of this project.

5.2. Product Perspective

The " FeelSpeak : Generating Emotional Speech " system is a software product that is intended to be used as a tool for generating emotional speech from text. The system is designed to be integrated into existing text-to-speech systems or used as a standalone tool for generating speech. The product is designed to provide users with a more engaging and emotionally expressive experience when interacting with speech-enabled systems, such as virtual assistants or customer service chatbots, text editors, etc…

The product is being developed with specialized techniques of natural language processing and speech technology. The system utilizes advanced machine learning algorithms to analyze text and generate emotional speech in real-time. The system is being developed with the goal of improving the user experience and increasing the effectiveness of speech-enabled systems across a variety of industries, including healthcare, customer service, and entertainment.

5.2.1 Product Features

The major features of the product include:

1. Text-to-Speech Conversion: The system can convert written text into emotional speech using various emotions such as happiness, sadness, anger, etc.
2. Emotion Selection: The user can select the emotion they want the speech to convey using a dropdown menu or other interface options.
3. Voice Selection: The user can select the voice they want the speech to be spoken in using a dropdown menu or other interface options.
4. Audio File Export: The system allows the user to export the generated speech as an audio file in popular formats such as MP3, WAV, etc.
5. Text Input: The user can input the text they want to be converted into emotional speech using a text box or other interface options.
6. Emotion Detection: The system can detect the emotion in the input text and suggest appropriate emotions for the speech output.
7. Multi-language Support: The system supports multiple languages for input text and speech output.
8. User Management: The system allows for user management features such as user registration, login, and profile management in the text editor.

9. Security and Privacy: The system implements security and privacy features to ensure user data and interactions are protected.
10. API Integration: The system can integrate with third-party APIs for additional features and functionalities.

5.2.2 Operating Environment

The operating environment for the system includes the following components:

- Hardware platform: The system can be run on any standard computer hardware with a processor, memory, and storage space capable of running the required software.
- Operating system: The system will be developed and tested on Windows and Linux operating systems. It will be compatible with Windows 7 or later, and Linux distributions such as Ubuntu, Fedora, and CentOS.
- Software components: The system requires the following software components to be installed:

Python 3.6 or later, TensorFlow 2.0 or later, NumPy, Pandas, NLTK, Flask

In addition, the system requires an internet connection to access cloud-based speech synthesis services.

5.2.3. General Constraints, Assumptions and Dependencies

The following are the general constraints, assumptions, and dependencies for the system:

- Hardware Limitations: The system requires a computer with a minimum configuration of a quad core processor, 8 Gigabyte RAM, and 1 Gigabyte free disk space.

- Operating System: The system will be developed and tested on Windows 10 (64-bit) operating system.

- Robust Dataset : This project requires a dataset which has labeled text with emotion and a corresponding speech dataset with emotion. Also, an unlabeled text dataset with emotion is required for validations.

- Assumptions: The system assumes that the input text is in the English language and contains no grammatical errors.

- Interface Dependencies: The system will use external libraries for speech synthesis and natural language processing.

- Safety and Security Considerations: The system should follow all the safety and security considerations related to data privacy and security.

Other assumptions are :

- Text entered will be in English Language with a UK/US accent.
- A total of five emotions are proposed to be detected, which include happy, sad, neutral, angry, sarcastic.
- Text is assumed to be an annotated form of text, i.e, conversational format of text, role based dialogue format, etc.
- The number of characters will be limited to a few hundreds at the beginning which can be scaled up later.
- Expectation of some words in the sentence that actually gives the emotion, otherwise, just specifying exclamation marks, will give neutral speech.
- The output of speech is only in a single voice. Multiple voices and tunes can be added later.

Dependencies:

1. Availability of labeled speech and text dataset: The availability of labeled speech and text data is a critical dependency for the project. Without this data, it is impossible to train and test the emotion recognition and conversion models.

2. Accuracy of the emotion recognition model: The accuracy of the emotion recognition model is a key dependency for the project. If the emotion recognition model is not accurate, it will produce incorrect emotional speech outputs, leading to a poor user experience.

3. Availability of processing power and resources: The availability of processing power and resources is also a significant dependency for the project. Without these resources, it is not possible to train and test the machine learning models effectively.

Impact of dependencies:

The impact of these dependencies is that they affect the project's timeline and overall success. Without adequate resources, and accurate models, the project may experience delays, quality issues, and poor user adoption. It is essential to manage these dependencies carefully and plan for contingencies to ensure project success.

5.2.4. Risks

As with any software development project, there are risks involved with the development of the proposed system. The following risks have been identified:

- Technical Risks:The complex nature of the underlying natural language processing algorithms poses a danger that the system may not be able to produce appropriate emotional speech from text. The risk can be reduced by thoroughly evaluating and enhancing the employed algorithms.

- Schedule Risks: There is a possibility of project delays owing to unexpected technical challenges or changes in project requirements. Setting realistic timeframes, routinely checking work, and having contingency plans in place might help mitigate this risk.

- Resource Risks: There is a risk of inadequate resources (such as hardware, software, or personnel) that may impact the project's success. This risk can be mitigated by careful planning, as well as having backup resources in case of emergencies.

- Security Risks:The system might be subject to security threats such as data leaks or hacking attempts. This threat can be reduced by establishing robust security measures like encryption and access controls, as well as monitoring the system on a frequent basis for any unusual activities.

5.3. Functional Requirements

5.3.1 Validity tests on inputs:

The system shall validate the input text to ensure that it is in the English language.

The system shall check the length of the input text to ensure that it is within the acceptable limits for the model. The system shall verify that the input text does not contain any profanity or offensive language.

Sequence of operations:

- The system shall first preprocess the input text to remove any unnecessary characters, punctuations, or digits.
- The system shall then tokenize the preprocessed text into words and phrases.
- The system shall apply the emotional embedding algorithm to each token to generate emotional features.
- The system shall use the emotional features to generate emotional speech signals using the TTS system.

5.3.2 Error handling and recovery:

The system shall generate an error message if the input text is not in English.

The system shall provide an error message if the input text is too long.

The system shall provide an error message if the emotional embedding algorithm fails to generate emotional features for a token. The system shall provide a fallback option to use a default emotion or neutral emotion if emotional speech signals cannot be generated for a given input.

5.3.3 Consequences of parameters:

The system shall adjust the emotional intensity of the speech signal based on the specified emotional parameters (e.g., happy, sad, angry, etc.).

The system shall vary the speaking rate of the generated speech signal based on the emotional parameters.

5.3.4 Relationship of outputs to inputs:

The system shall generate emotional speech signals that match the emotional content of the input text. The system shall provide an option to output the emotional speech signal in different audio formats.

5.4. External Interface Requirements

5.4.1. User Interfaces

The system will have a user interface that allows users to input the text for which they want to generate emotional speech. The user interface will consist of a text box where users can enter the text and a button to initiate the speech generation process. The user interface will follow standard GUI design principles, including consistent layouts, color schemes, and font sizes.

The system will display the generated speech output to the user in the form of an audio file. The audio file will be played back using the user's default audio player, and the system will not provide any playback controls or options. The audio file will be available for download to the user's device.

In case of errors, the system will display appropriate error messages to the user in a separate pop-up window. The error messages will provide detailed information about the error,

including possible solutions or workarounds. The system will also provide a help button that users can click to access a user manual or other resources.

5.4.2. Hardware Requirements

The hardware requirements for the "Generation of emotional speech from text" system are as follows:

Computer or mobile device with at least 8GB of RAM and a dual-core processor.
Operating System: Windows 10 or later, MacOS, or Linux.
Sound card and microphone for recording audio.
Speakers or headphones for audio playback.
Internet connection for downloading and installing required software packages.

The system should be compatible with a range of hardware devices, including laptops, desktops, and mobile devices. The system should support standard audio input/output devices and protocols, such as 3.5mm audio jacks, USB, and Bluetooth. The system should also be capable of using the internet connection to access and use cloud-based resources if required.

5.4.3. Software Requirements

Since the project is about developing a software system, there are no software requirements beyond the system itself. However, it can be assumed that the software requirements for this system include:

• Name and Description: The software system for generating emotional speech from text

• Version / Release Number: Initial release version 1.0

• Databases: No specific database requirements have been mentioned

• Operating Systems: The system should be able to execute the program with major OS like macos, linux and windows.

• Tools and libraries: The system may require various tools and libraries to generate speech from text, such as text-to-speech (TTS) engines, natural language processing (NLP) libraries, and audio processing libraries.

• Source (if any): The source code for the software system may be provided to the client or maintained by the development team, depending on the agreement between the parties involved.

5.4.4. Communication Interfaces

- Network Interface: If the system is designed to work in a client-server architecture, it should have an interface to communicate with the server over a network. This interface must support standard network protocols, such as TCP/IP or HTTP.

- Text Input/Output Interface: The system must have an interface to capture text input from the user and generate emotional speech output based on the input text. The interface must support standard text formats, such as plain text.

- External API Interface: The system may need to integrate with external APIs, such as speech recognition or natural language processing APIs. The interface must be compatible with the API protocols and specifications. This is seen in the future maintenance of the project.

5.5. Non-Functional Requirements

5.5.1. Performance Requirement

- The system shall generate emotional speech from text within a few seconds of receiving the input. This is because the input size may vary.
- The system shall be available for use 24/7 with an uptime of at least 99%.
- The system shall have a maximum error rate of 15 % in generating emotional speech from text.
- The system shall be able to process text inputs of up to 1000 characters in length with the minimum being 150 characters. This can be slowly scaled up in future.

In terms of quality attributes, some possible non-functional requirements could include:

- The system shall be robust and able to handle unexpected inputs and errors without crashing or corrupting data. We shall include various error handling cases
- The system shall have a user-friendly interface, basically a text editor with clear and concise error messages.
- The system shall be secure, with all user data and interactions encrypted and stored in a secure database.
- The model will have user feedback regarding the correct emotions to ensure that model will learn continuously and give a robust performance.

5.5.2. Safety Requirements

As the product involves generation of emotional speech from text, there are no safety requirements to be addressed in the product. Of course, the text entered by the user will be safe in one's system.

5.5.3. Security Requirements

- User Authentication

The system must require users to provide valid login credentials before allowing access to the system. Passwords must meet complexity requirements and be stored securely.

- Authorization

The system must enforce role-based access control, that is the level of hierarchy to ensure that users can only access functionality and data that is relevant to their role. This kind of role-based access is very seldom to appear in this application.

- Data Privacy

The system must comply with applicable data privacy regulations and protect sensitive data from unauthorized access. Any personal or sensitive data collected by the system must be encrypted in transit and at rest.

- Security Auditing

The system must maintain an audit trail of all user actions, including login attempts, data access, and any changes made to the system configuration. The audit trail must be tamper-evident and securely stored.

5.6. Other Requirements

- Ethical Considerations

The system must comply with ethical considerations in generating emotional speech. The system should not be used to generate hate speech or any other form of speech that may cause harm to others. It should also respect the privacy and consent of the user.

- Data Protection Requirements

The system must comply with data protection requirements. It should not collect or store any personal data of the user without their explicit consent. The system should also provide an option for users to delete their data.

- Accessibility Requirements

The system must adhere to these requirements in order to guarantee that it is usable by all users.specific standards for accessibility. These standards include the availability of text-to-speech or audio descriptions, which are crucial for users with disabilities.disabilities. Additionally, user-friendliness should be considered when designing the system to make sure that it is simple to use for all users.

- Compatibility Requirements

For the system to be accessible to a variety of users, it must be compatible with a variety of hardware and operating systems. It ought to work flawlessly on several platforms, including desktop and mobile gadgets.

6. SYSTEM DESIGN

6.1 Architecture Diagram

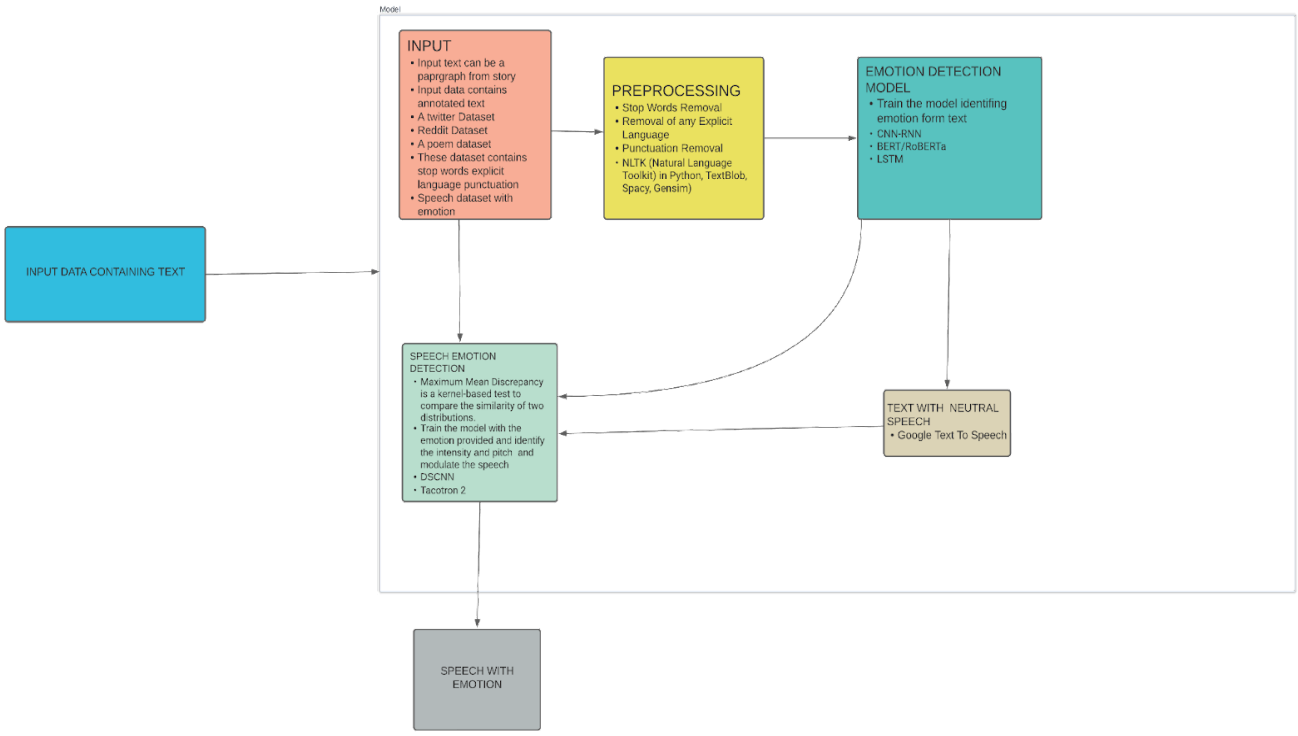

Figure 6.1

6.2 Flow/Sequence Diagram

6.2.1 Training

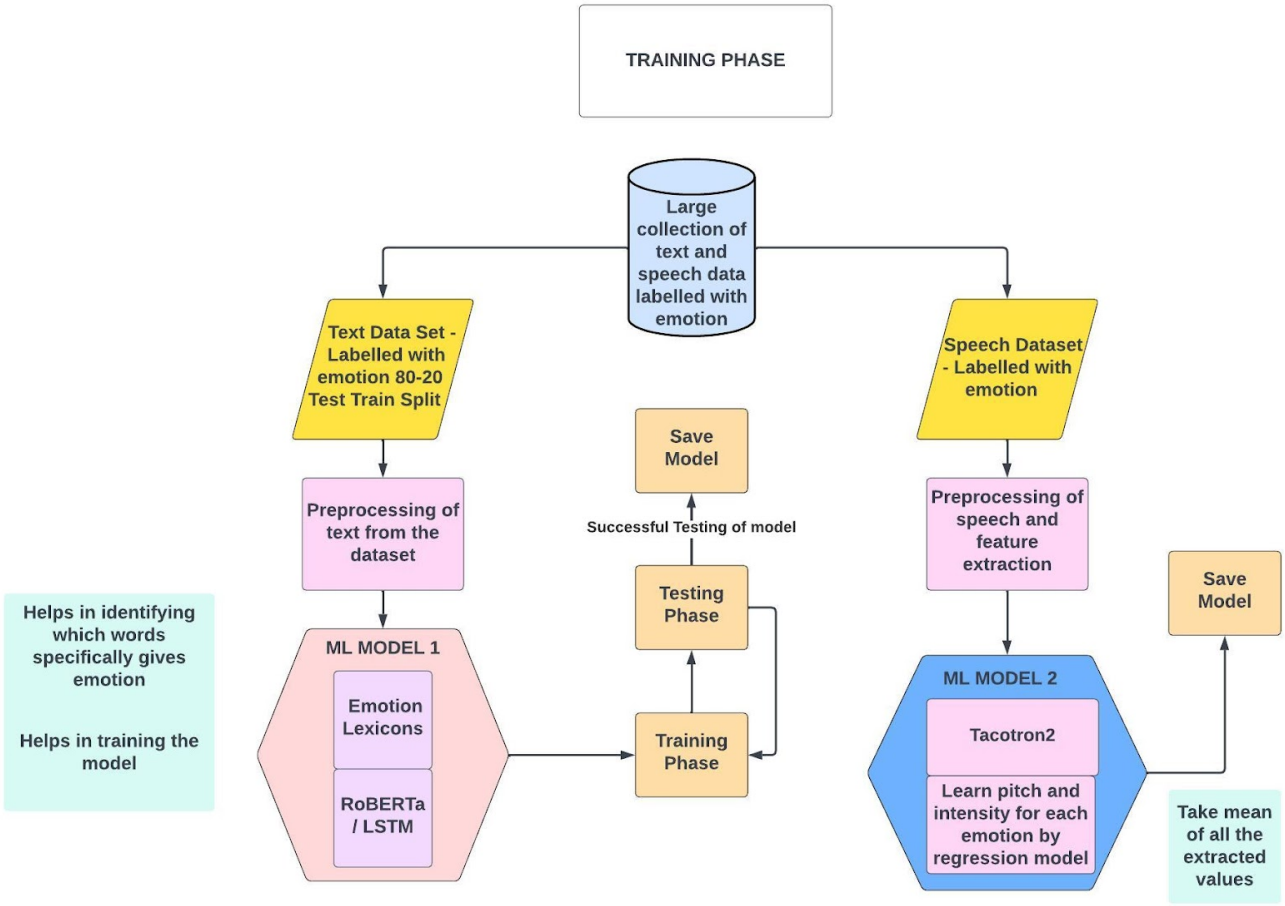

Figure 6.2.1

6.2.2 Testing

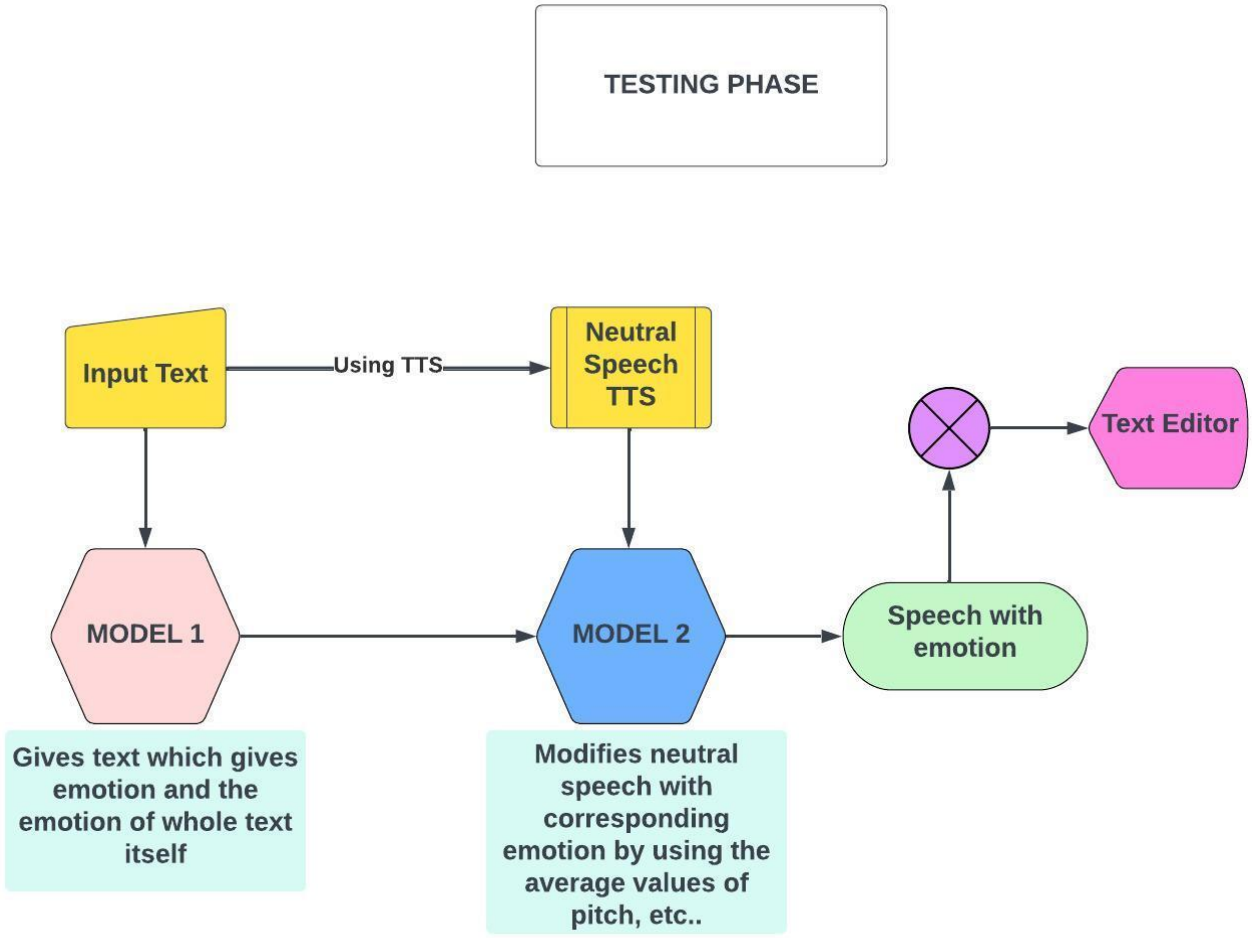

Figure 6.2.2

6.3 Swimlane Diagram

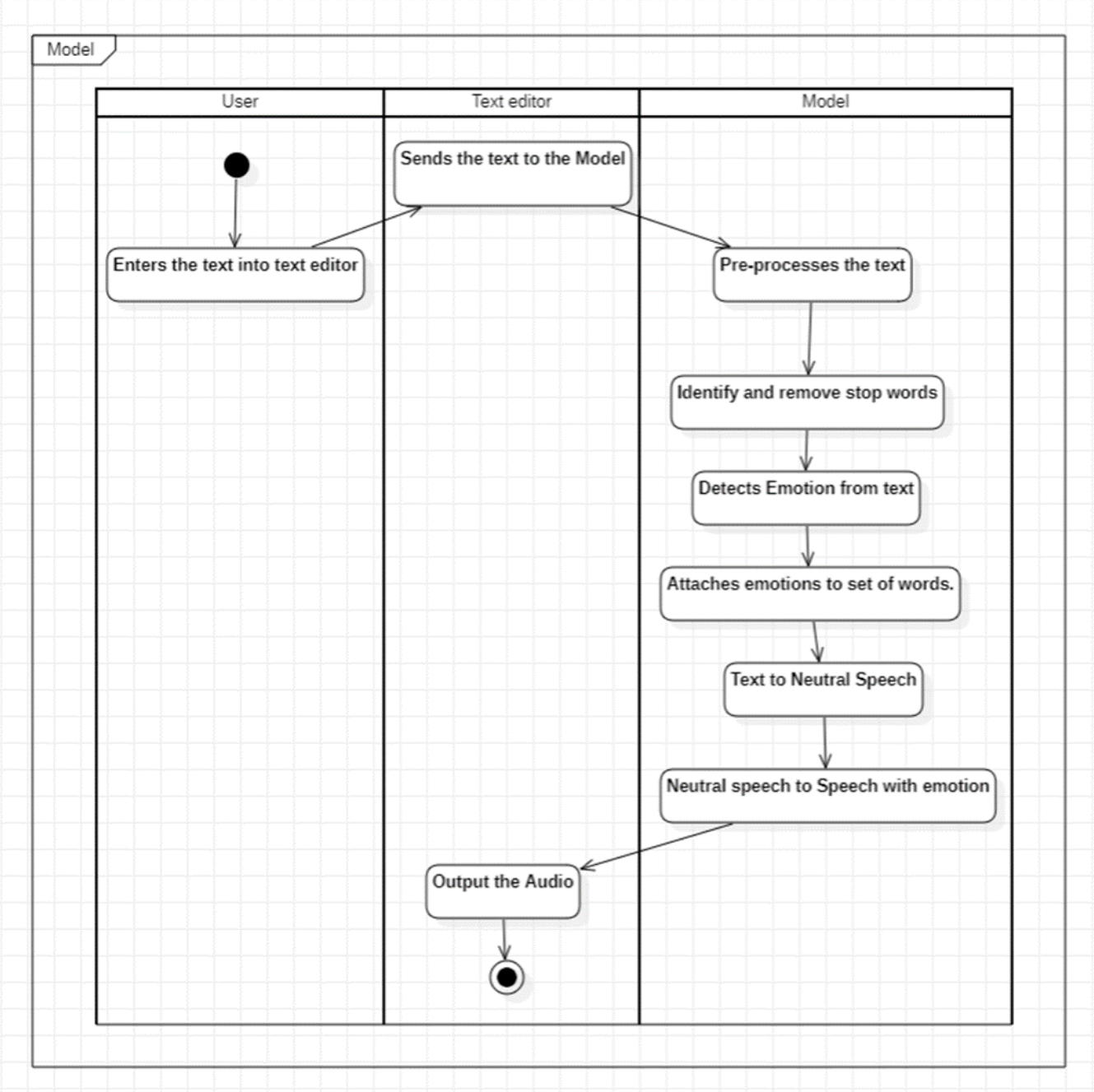

Figure 6.3

- Swimlane diagrams are flowcharts that represent the flow of processes across different departments, roles, or individuals in an organization. They use "lanes" or "swim lanes" to represent different parties involved in the process and can be used to map out complex processes, clarify responsibilities, identify inefficiencies, and improve communication and collaboration among different departments or teams.
- In the given diagram we can see how the control of the processes shifts from User lane to the Text Editor. The Text Editor shifts to the model lane and back to the Text Editor. The Model Converts the User's Input Text to Speech with emotion and gives out this audio output through the Text editor.

6.4 User Interface Diagrams

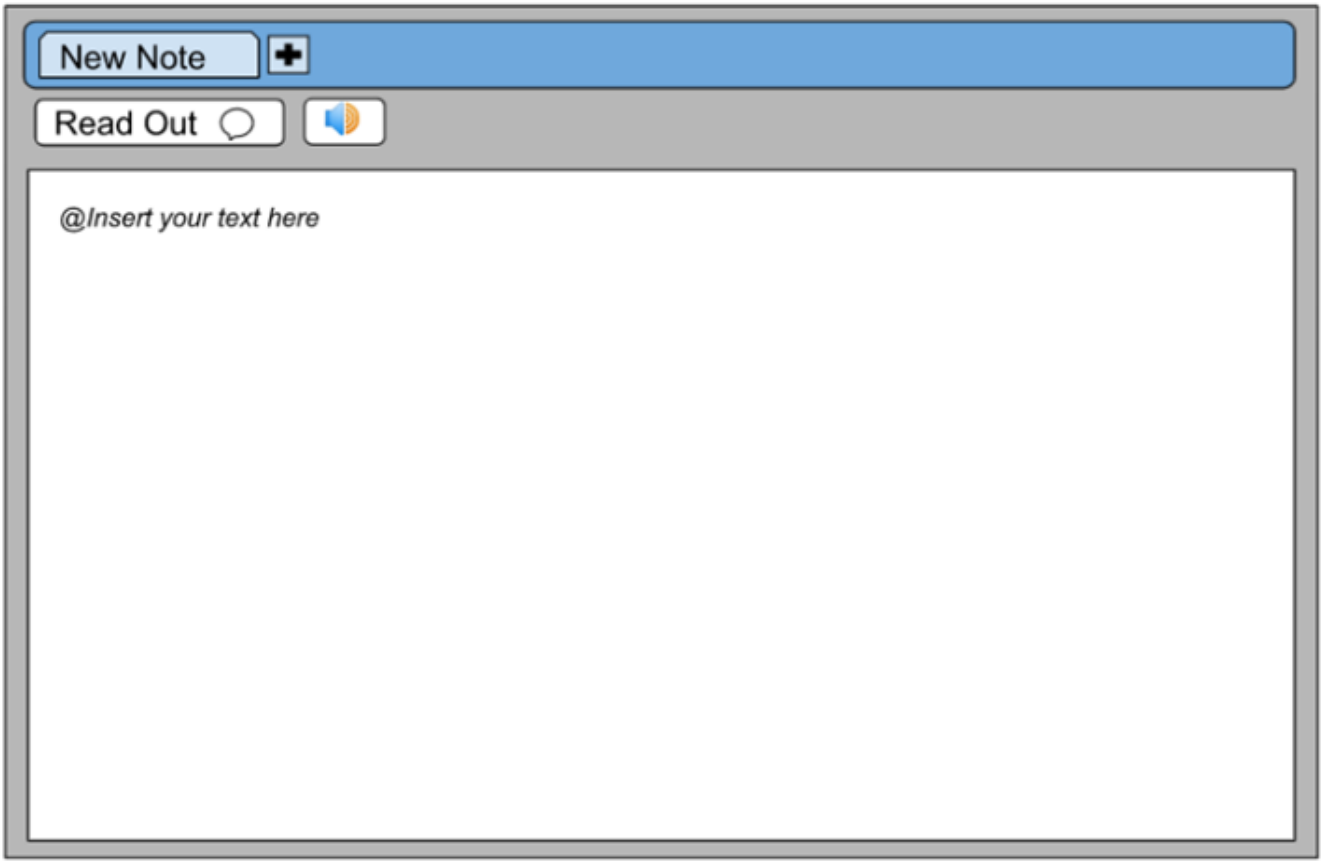

Figure 6.4

The Text Editor acts as the Interface between The User and the Model which gives out speech with emotion as output through the text editor.

6.5 External Interfaces

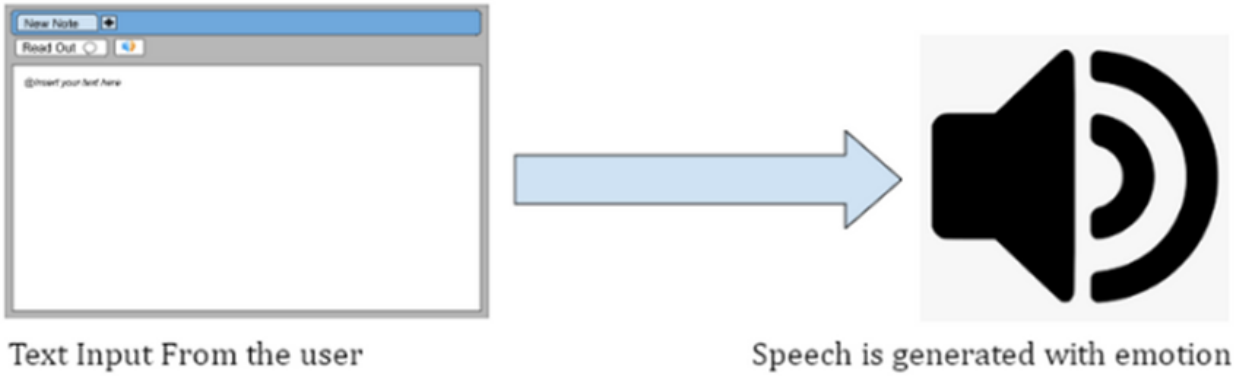

Figure 6.5

The Text Editor acts as the Interface between The User and the Model which gives out speech with emotion which is the output through the device's in-built speaker or an external speaker.

6.6 Packaging and Deployment Diagram

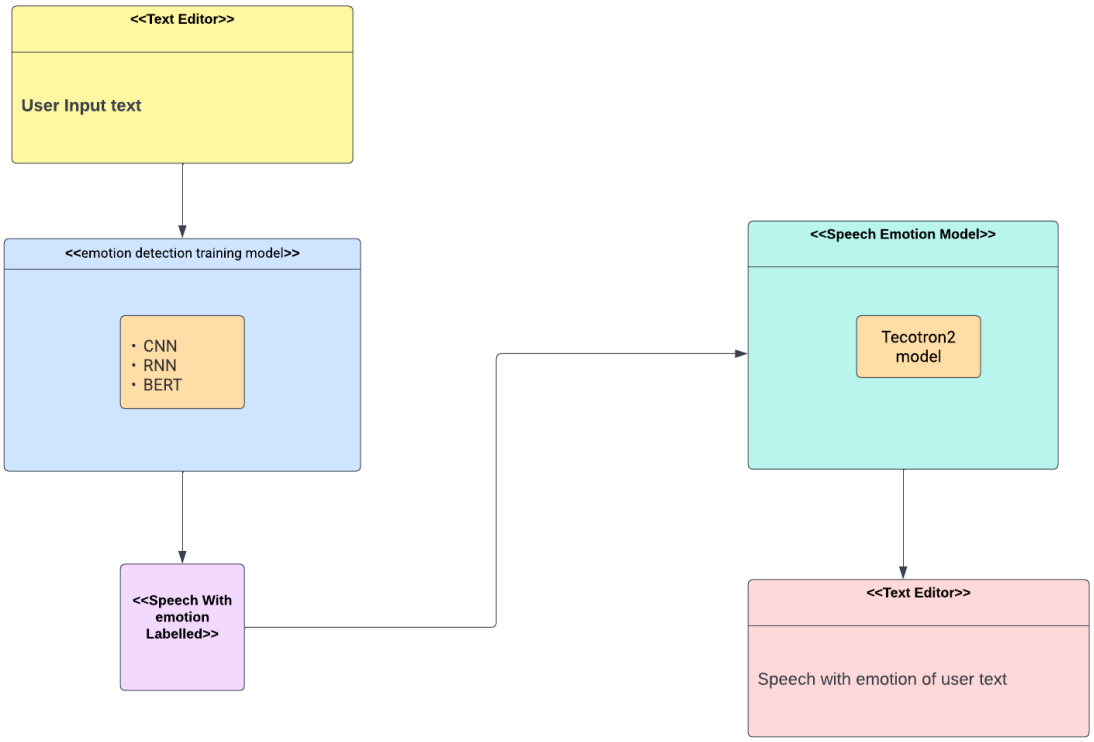

Figure 6.6

7. IMPLEMENTATION STEPS

7.1. Implementing Training Part

1. Collect a large and diverse dataset of text with labeled emotions. This dataset should include a variety of emotions such as happiness, sadness, anger, and fear. The dataset can be obtained from publicly available emotion-labeled datasets or can be created by manually labeling the emotions in text.

2. Preprocess the text data by cleaning and normalizing the text, removing stop words, and converting the text to a numerical representation such as word embeddings.

3. Split the dataset into 80% training and 20% validation sets.

4. Train an emotion detection model on the training set using a pre-trained language model such as BERT or GPT. The model should be fine-tuned on the labeled dataset to predict the emotion in text.

5. Evaluate the performance of the emotion detection model on the validation set using objective metrics such as accuracy, precision, recall, and F1-score.

6. If the performance of the model is satisfactory, save the model for future use.

7. Collect a large dataset of speech with labeled emotions. This dataset should include a variety of emotions such as happiness, sadness, anger, and fear. The dataset can be obtained from publicly available emotion-labeled datasets or can be created by recording speech with different emotions and labeling them.

8. Preprocess the speech data by extracting speech features such as pitch, intensity, and modulation.

9. Train a speech emotion recognition model on the training set using a pre-trained model such as SER-ResNet or SE-ResNeXt. The model should be fine-tuned on the labeled dataset to predict the emotion in speech. Later we can build our own model for the same and evaluate the results.

10. Evaluate the performance of the speech emotion recognition model on the validation set using objective metrics such as accuracy, precision, recall, and F1-score.

11. If the performance of the model is satisfactory, save the model for future use.

7.2 Implementing Testing Part

1. Preprocess the input text by cleaning and normalizing the text, removing stop words, and converting the text to a numerical representation such as word embeddings.

2. Pass the preprocessed text to the emotion detection model to predict the emotion in the text.

3. Generate neutral speech from the input text using a text-to-speech (TTS) system.

4. Retrieve the speech features (pitch, intensity, and modulation) for the predicted emotion from the speech emotion recognition model.

5. Apply the retrieved speech features to the neutral speech using a signal processing algorithm.

6. Play the generated speech with the predicted emotion.

7. Evaluate the generated speech using subjective metrics such as the Geneva Emotional Evaluation Scale (GEES) to measure the perceived emotion in the generated speech.

8. Iterate and refine the system as necessary based on the evaluation results.

7.3 Sarcasm

Data collection: Collect a dataset of text that includes both sarcastic and non-sarcastic examples. This dataset should be large enough to ensure the model can generalize well.

Data preprocessing: Preprocess the data by removing stop words, stemming, lemmatization, and converting text to lowercase.

Feature extraction: n-grams, TF-IDF, and word embeddings can be used to extract features from preprocessed data.

Model selection: Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Neural Networks can be used as models to detect sarcasm from text.

Training: We will be using different machine learning algorithms such as Neural Networks (NN), Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Naïve Bayes (NB) on the datasets.Adjust the model hyperparameters(like C and sigma used in SVM) to improve performance on the validation set.

Testing: To evaluate the model we will be using metrics such as F1-score, accuracy, precision and recall.

7.4 Tools used

Pandas, numpy, tensorflow keras, nltk , re, seaborn, matplotlib, sklearn and lime,Tacotron2, librosa, pyaudio, pydub, etc…

8. CONCLUSION OF CAPSTONE PROJECT PHASE-1

- As a part of capstone phase-1 literature survey and high level design documents have been created .
- Various datasets have been found out for training our model
- We also have a text editor ready-in-hand for the model to be integrated with.

9. PLAN OF WORK FOR CAPSTONE PROJECT PHASE-2

- Data Preprocessing: The team will preprocess the input dataset which has to be passed to our model.
- Model Training: Training the SER & TTS model which uses various machine learning algorithms.
- Evaluation and Testing: After the training is done the team will validate the model to check if the algorithm used in the model is giving a better accuracy . If not the team will decide to use a another model
- Refinement and Optimization: The team will try to optimize the parameters on the model to improve the accuracy
- Documentation and Reporting: Create a document for the model which would help in the understanding how the model works and a final report will be created to tell what scores the model has given.

REFERENCE/ BIBLIOGRAPHY

[1] X. Cai, D. Dai, Z. Wu, X. Li, J. Li and H. Meng, "Emotion Controllable Speech Synthesis Using Emotion-Unlabeled Dataset with the Assistance of Cross-Domain Speech Emotion Recognition," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 5734-5738, doi: 10.1109/ICASSP39728.2021.9413907

[2] Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data Saurav Pradha School of Computing and Mathematics Charles Sturt University Melbourne, Victoria, Senior Australia saurav.pradha54@gmail.com Malka N. Halgamuge Member, IEEE Dep. of Electrical and Electronic Engineering The University of Melbourne Victoria 3010, Australia malka.nisha@unimelb.edu.au Nguyen Tran Quoc Vinh Faculty of Information Technology The University of Da Nang - University of Science and Education, Vietnam ntquocvinh@ued.udn.vn.

[3] Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, 2018Navdeep Jaitly, Zongheng Yang, Ying Xiao, Z. Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Robert A. J. Clark and Rif A. Saurous. “Tacotron: Towards End-to- End Speech Synthesis.” Interspeech (2017).

[4] P. Chandra et al., "Contextual Emotion Detection in Text using Deep Learning and Big Data," 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2022, pp. 1- 5, doi: 10.1109/ICCSEA54677.2022.9936154.

[5] Ren, Yi, Ruan, Yangjun, Tan, Xu, Qin, Tao, Zhao, Sheng, Zhao, Zhou, and Tie Liu. "FastSpeech: Fast, Robust and Controllable Text to Speech." ArXiv, (2019). Accessed February 10, 2023. https://doi.org/10.48550/arXiv.1905.09263.

[6] Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J. (2016) Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. Proc. 9th ISCA Speech Synthesis Workshop, 146-152.

[7] S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," in IEEE Transactions on Speech and Audio Processing, vol. 12, no. 4, pp. 401-408, July 2004

[8] Text to Speech Conversion with Emotion Detection Article in International Journal of Applied Engineering Research · January 2018 Srinivasan Rajendran SRM Institute of Science and Technology 19 PUBLICATIONS 66 CITATIONS

[9] A Comprehensive Review of Speech Emotion
Recognition SystemsTAIBA MAJID WANI 1, TEDDY SURYA GUNAWAN 1,3, (Senior Member, IEEE), SYED ASIF AHMAD QADRI 1, MIRA KARTIWI 2, (Member, IEEE), AND ELIATHAMBY AMBIKAIRAJAH 3, (Senior Member, IEEE)

[10] SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORK CONSIDERING VERBAL AND NONVERBAL
SPEECH SOUNDS Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen, Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

[11] Carrillo-de-Albornoz, Jorge & Plaza, Laura & Gervás, Pablo. (2012). SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis.

[12] Huan, Jeow & Sekh, Arif Ahmed & Quek, Chai & Prasad, Dilip. (2022). Emotionally charged text classification with deep learning and sentiment semantic. Neural Computing and Applications. 34. 10.1007/s00521-021-06542-1.

[13] U. Rashid, M. W. Iqbal, M. A. Skiandar, M. Q. Raiz, M. R. Naqvi and S. K. Shahzad, "Emotion Detection of Contextual Text using Deep learning," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 2020, pp. 1-5, doi: 10.1109/ISMSIT50672.2020.9255279.

[14] A Comprehensive Review of Speech Emotion
Published in: IEEE Access (Volume: 9)
Page(s): 47795 - 47814
Date of Publication: 22 March 2021
Electronic ISSN: 2169-3536
INSPEC Accession Number: 20965838
DOI: 10.1109/ACCESS.2021.3068045
Publisher: IEEE

[15] Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional
Neural Networks - 2020
Published in: 2020 6th International Conference on Wireless and Telematics (ICWT)
Date of Conference: 03-04 September 2020

Date Added to IEEE Xplore: 03 November 2020
ISBN Information:
INSPEC Accession Number: 20133021
DOI: 10.1109/ICWT50448.2020.9243622

[16] Demszky, Dorottya, et al. "GoEmotions: A dataset of fine-grained emotions." arXiv preprint
arXiv:2005.00547 (2020)

[17] Text to Speech Conversion with Emotion Detection Article in International Journal of Applied Engineering
Research · January 2018

Voice_datasets

APPENDIX A : DEFINITIONS, ACRONYMS AND ABBREVIATIONS

Emotion detection model: A model that analyzes text or speech and identifies the corresponding emotion, such as anger, sadness, happiness, or sarcasm

Preprocessing: The process of cleaning and preparing input data for further processing.

Stop words: Commonly used words that are often removed from text data during preprocessing."

Explicit words: Words that may be inappropriate or offensive in certain contexts.

Punctuation: Marks such as periods, commas, and quotation marks that are used to separate and structure text.

Deep learning: A subset of machine learning that involves training neural networks to perform complex tasks.

Sarcasm: A form of verbal irony in which the intended meaning of a word or phrase is the opposite of its literal or expected meaning.

UK/US accent: A way of pronouncing English that is specific to either the United Kingdom or the United States.

Annotated text: Text that includes additional information, such as labels or tags, to provide context and meaning.

Role-based dialogue format: A format of text that is used in a dialogue between two or more people, with each person assigned a specific role.

GPU: Graphics Processing Unit, a type of computer processor that is optimized for parallel processing and commonly used for machine learning applications.

Prosody: The patterns of stress and intonation in speech that convey emotional content.

NLP - Natural Language Processing

LSTM - Long Short-Term Memory

CNN - Convolutional Neural Network

RNN - Recurrent Neural Network

POS - Part-of-Speech

SVD - Singular Value Decomposition

PCA - Principal Component Analysis

GUI - Graphical User Interface

UI - User Interface

MVP - Minimum Viable Product

PLAGIARISM REPORT

REPORT_CS235_320_355_362

ORIGINALITY REPORT

5%	3%	4%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	1%
2	dspace.lib.cranfield.ac.uk Internet Source	1%
3	export.arxiv.org Internet Source	<1%
4	icwt-seei.org Internet Source	<1%
5	"Recent Developments in Electronics and Communication Systems", IOS Press, 2023 Publication	<1%
6	Jeow Li Huan, Arif Ahmed Sekh, Chai Quek, Dilip K. Prasad. "Emotionally charged text classification with deep learning and sentiment semantic", Neural Computing and Applications, 2021 Publication	<1%
7	munin.uit.no Internet Source	<1%