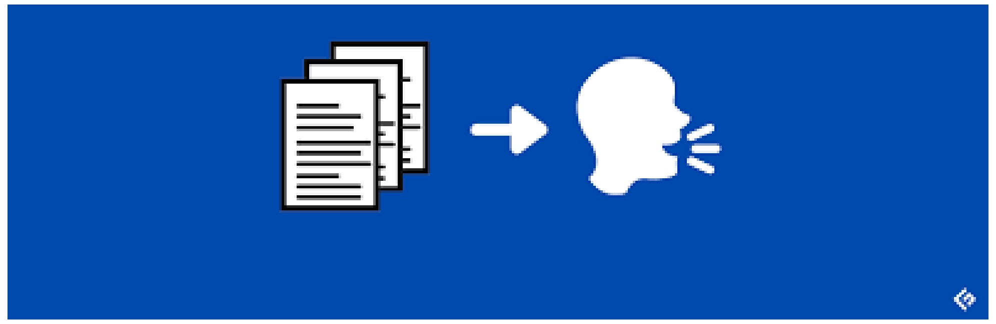
UE20CS390A – Capstone Project Phase – 1

SEMESTER - VI

END SEMESTER ASSESSMENT



Project Title : FeelSpeak: Generating Emotional Speech with Deep Learning
Project ID : PW23_VRB_07
Project Guide : Prof. V R Badri Prasad
Project Team : 235_320_345_362

Outline

- Problem Statement
- Abstract and Scope
- Literature Survey
- Suggestions from Review – 3
- Design Approach
- Design Constraints, Assumptions & Dependencies
- Proposed Methodology / Approach
- Architecture
- Design Description
- Technologies Used
- Project Progress
- References

Abstract and Scope

Abstract:

- This project aims to develop a system that can generate emotional speech from given input text.
- The system will identify the emotions expressed in the text and generate speech with appropriate prosodic features to convey those emotions effectively.
- The project involves natural language processing, speech synthesis, and emotion recognition tasks.

Scope:

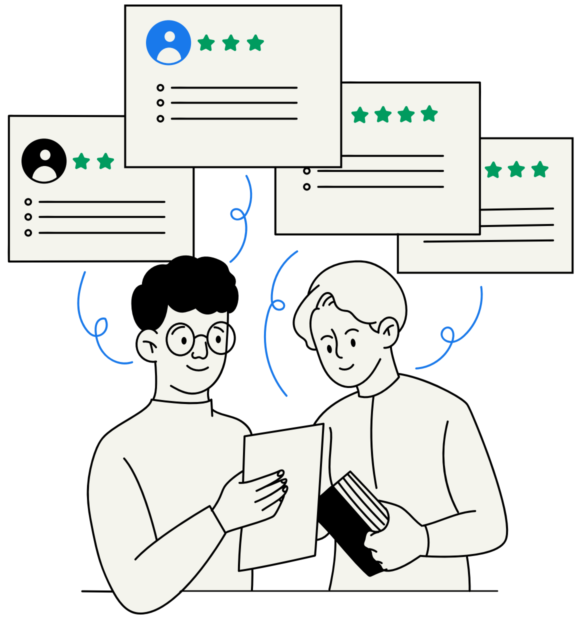
- The scope of this project includes developing a system that can accurately identify the emotional content of given input text and generate speech with appropriate prosodic features to reflect those emotions.
- The system will involve various tasks such as NLP and text emotion detection, speech synthesis, and emotion recognition.
- The NLP and text emotion detection component will parse the input text to identify its structure, meaning, and emotional content.
- The speech synthesis component will generate speech that accurately reflects the emotions expressed in the input text, and the emotion recognition component will map the emotional content of the input text to appropriate prosodic features.

Problem Statement

- The current state-of-the-art text-to-speech systems do not adequately convey emotional content, which is crucial for effective communication.
- This project aims to address this limitation by developing a system that can generate emotional speech from given input text.
- The challenge is to accurately identify the emotions expressed in the text and generate speech with appropriate prosodic features to convey those emotions effectively.
- This project involves various technical challenges such as NLP, text emotion detection, speech synthesis, and emotion recognition, and its success will greatly improve the quality of synthesized speech for various applications.

Suggestions from Review - 3

- As an update on the project progress, the team has successfully updated the flow diagram to reflect any changes or modifications to the project plan that have been made since the initial design phase.
- Additionally, the team has found suitable dataset with more sentences to use in training the emotion recognition and TTS models.



Literature Survey-ROHIT

Introduction:

- The aim of this paper is to develop an emotion controllable speech synthesis system using an emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition.
- The proposed system uses a text-to-speech synthesis model to generate speech from input text and a speech emotion recognition model to recognize the emotion in the generated speech.
- By using an emotion-unlabeled dataset, the proposed system is able to generate speech with a wide range of emotions.

Models used:

- SER Model: SER model consists of a feature extraction encoder and an emotion classifier. The encoder is a CNN-RNN network.
- Cross-domain training :TTS and SER datasets are quite different in speakers,recording devices and recording environment,considering that the training of MMD is stable and our TTS dataset has no available emotion labels for parameter tuning,we choose the MMD method for our cross-domain SER task
- GST-based emotional TTS model:Emotional TTS model consists of a Reference encoder and GST module

Literature Survey-ROHIT

Technology used:

- CNN-RNN:It is used for feature extraction in the emotion recognition model. The network takes the log mel spectrum as input feature and outputs a feature vector for the emotion classification task.
- Tacotron2:It is a deep neural network architecture used for text-to-speech (TTS) synthesis

Pros:

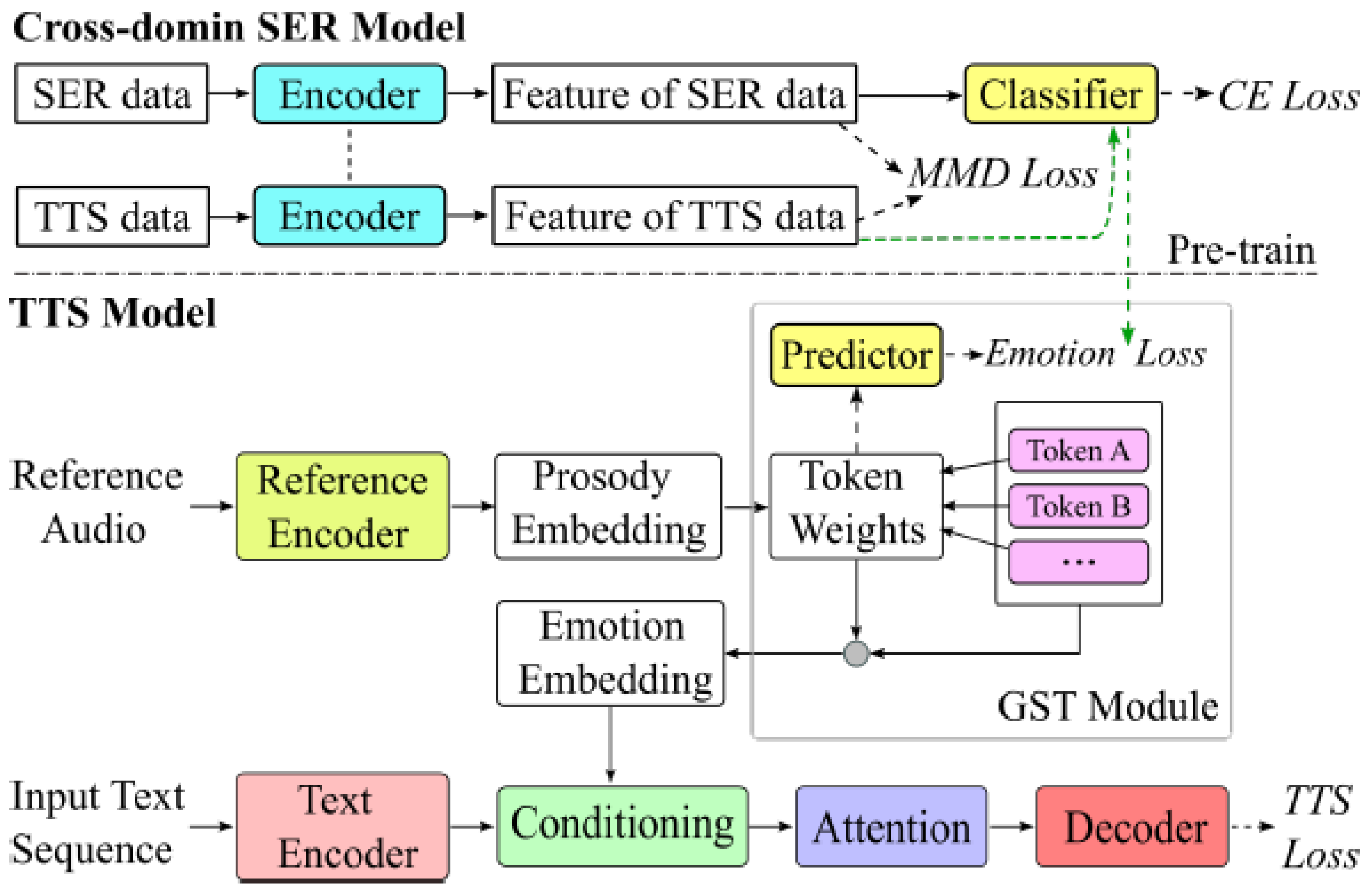
- The emotional TTS system proposed in the paper achieves high-quality speech synthesis with emotional content, which can enhance the naturalness and expressiveness of synthesized speech.
- The cross-domain training approach enables the emotional TTS model to be trained on a limited emotional speech dataset by leveraging a large-scale neutral speech dataset. This could be useful for scenarios where limited emotional speech data is available.

Cons:

- Computational complexity: The emotional TTS model has a relatively large number of parameters, which may make it computationally expensive to train and use in real-time applications.

Literature Survey-ROHIT

Result: 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.



Literature Survey - S M SUTHARSAN RAJ

PAPER : Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks - 2020

Introduction:

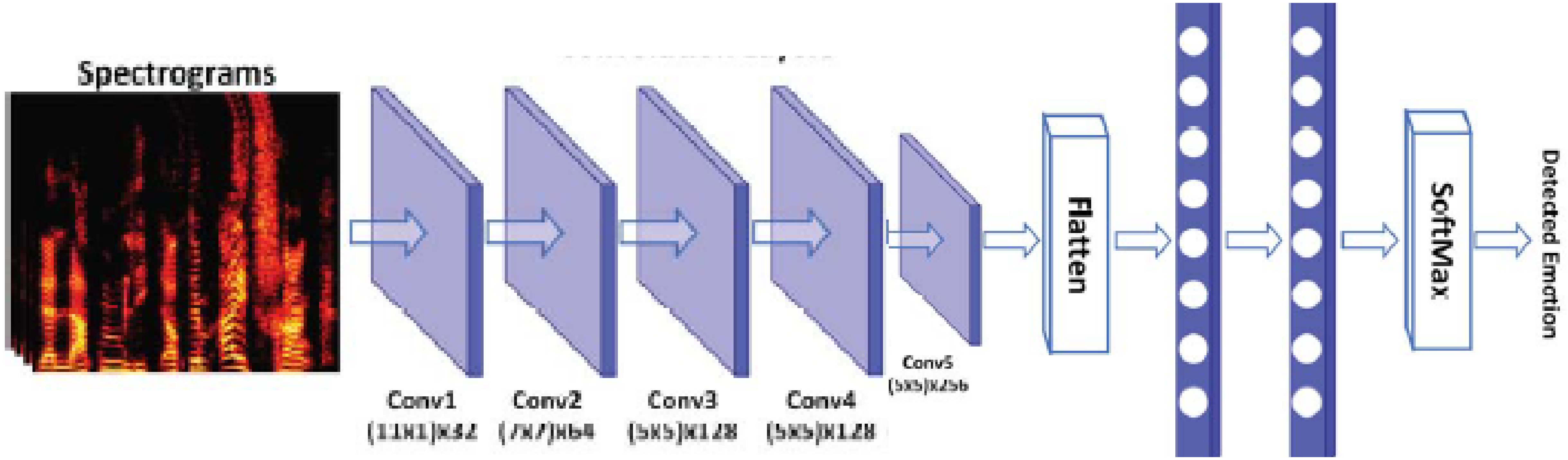
- This paper proposes a speech emotion recognition system using Convolutional Neural Networks (CNNs) and Deep Stride Convolutional Neural Networks (DSCNNs). The authors argue that these models can improve the accuracy of speech emotion recognition, especially for complex emotions.
- The literature review section of the paper discusses various approaches to speech emotion recognition, including traditional feature-based methods and deep learning-based methods. The authors note that deep learning-based methods have shown significant improvement in recent years due to their ability to learn features directly from raw data.

Methodology :

- The paper describes the proposed methodology, which involves preprocessing the speech signal, extracting features using Mel-frequency cepstral coefficients (MFCCs), and training CNNs and DSCNNs for emotion classification.
- The authors compare the performance of the proposed models with other deep learning-based models and traditional feature-based models on two publicly available datasets (IEMOCAP and Berlin Emotional Speech Database). The experimental results show that the proposed models outperform other models and achieve state-of-the-art performance in speech emotion recognition.

Literature Survey - S M SUTHARSAN RAJ

- DSCNN showed a great advantage of feature extraction of input speech signals. Here's how the author approached it :
- 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.



Literature Survey - S M SUTHARSAN RAJ

Results, accuracy and Conclusion

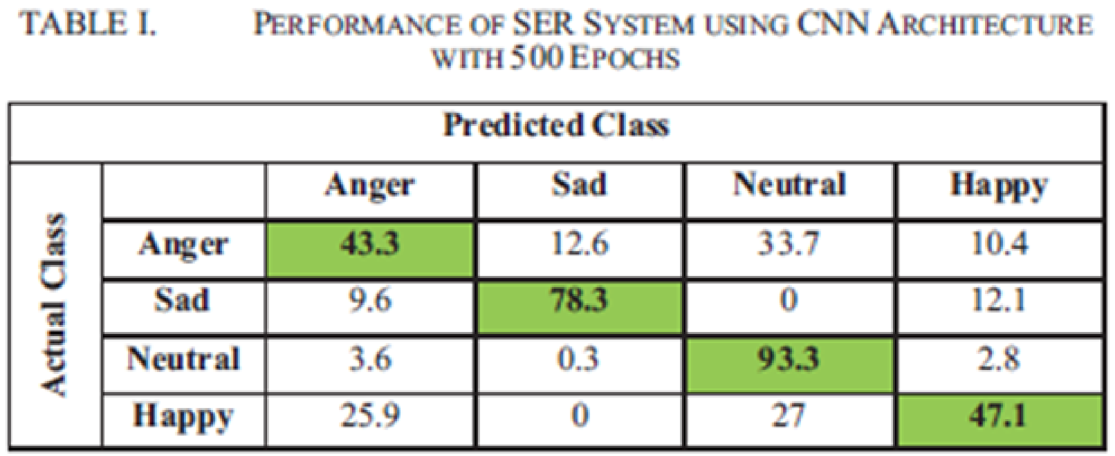


TABLE I. PERFORMANCE OF SER SYSTEM USING CNN ARCHITECTURE WITH 500 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	43.3	12.6	33.7	10.4
	Sad	9.6	78.3	0	12.1
	Neutral	3.6	0.3	93.3	2.8
	Happy	25.9	0	27	47.1

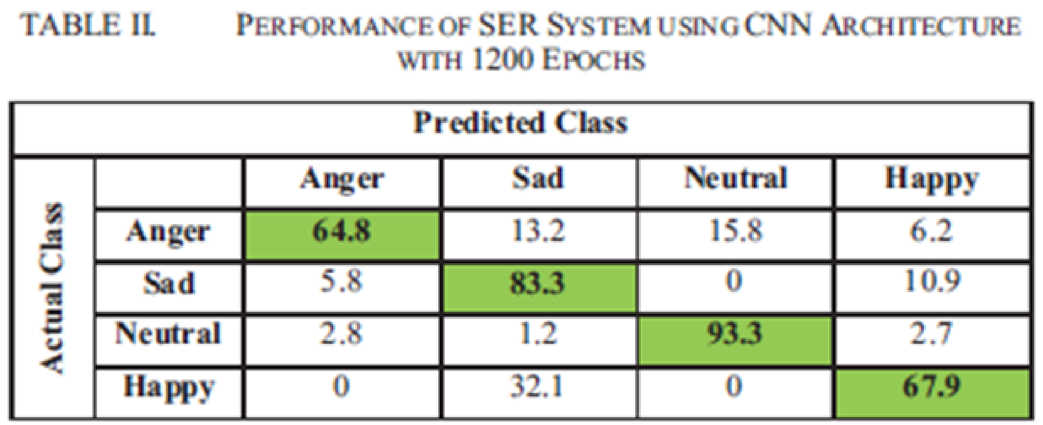


TABLE II. PERFORMANCE OF SER SYSTEM USING CNN ARCHITECTURE WITH 1200 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	64.8	13.2	15.8	6.2
	Sad	5.8	83.3	0	10.9
	Neutral	2.8	1.2	93.3	2.7
	Happy	0	32.1	0	67.9

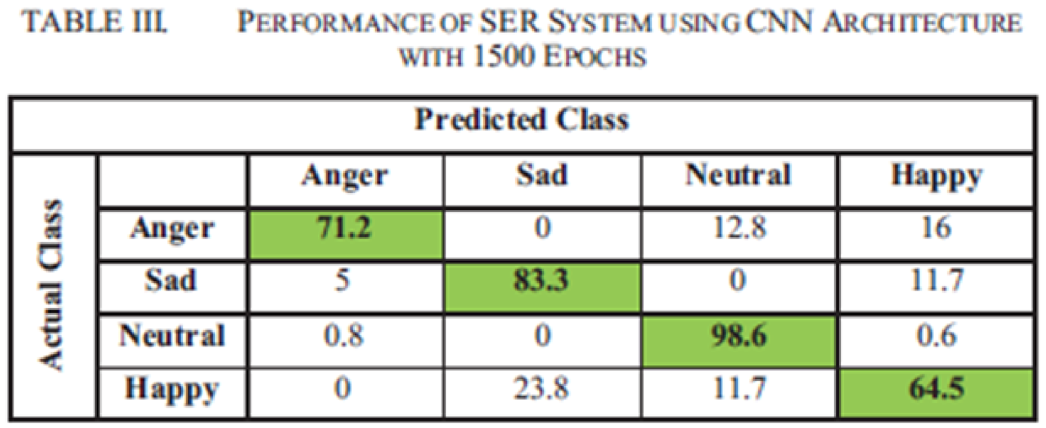


TABLE III. PERFORMANCE OF SER SYSTEM USING CNN ARCHITECTURE WITH 1500 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	71.2	0	12.8	16
	Sad	5	83.3	0	11.7
	Neutral	0.8	0	98.6	0.6
	Happy	0	23.8	11.7	64.5

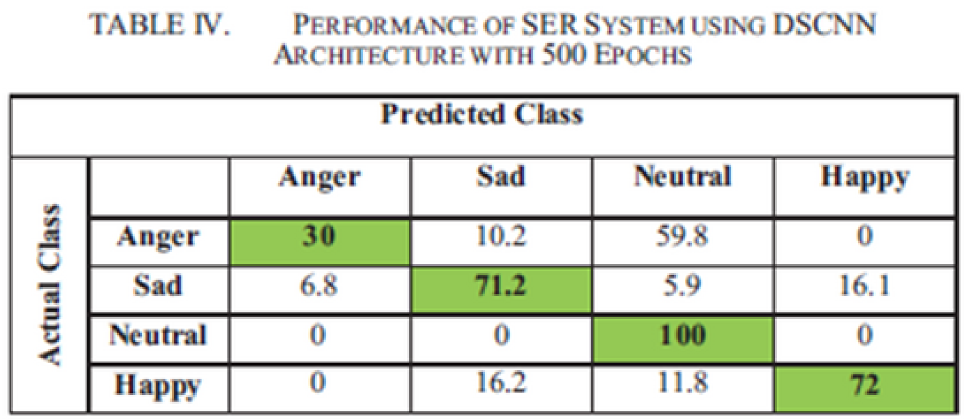


TABLE IV. PERFORMANCE OF SER SYSTEM USING DSCNN ARCHITECTURE WITH 500 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	30	10.2	59.8	0
	Sad	6.8	71.2	5.9	16.1
	Neutral	0	0	100	0
	Happy	0	16.2	11.8	72

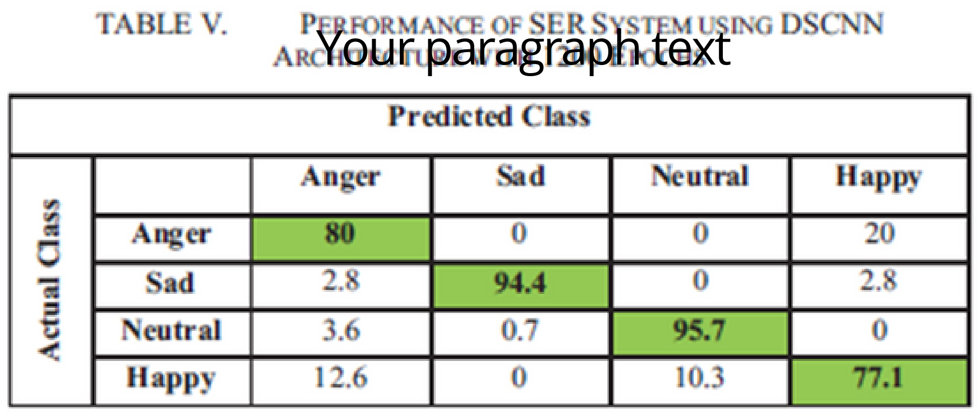


TABLE V. PERFORMANCE OF SER SYSTEM USING DSCNN ARCHITECTURE WITH 1200 EPOCHS

Your paragraph text

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	80	0	0	20
	Sad	2.8	94.4	0	2.8
	Neutral	3.6	0.7	95.7	0
	Happy	12.6	0	10.3	77.1

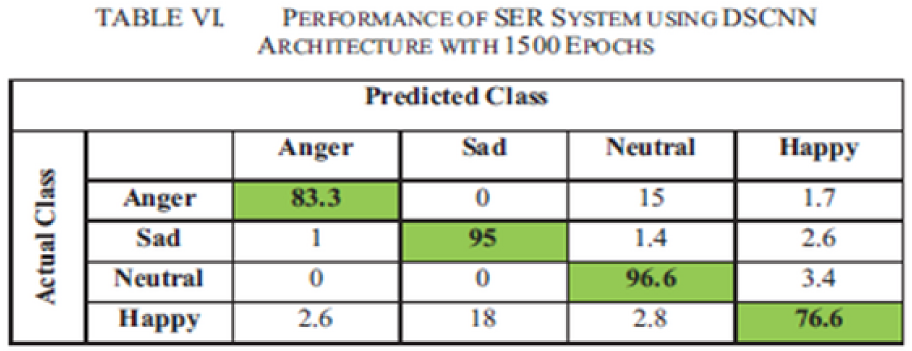


TABLE VI. PERFORMANCE OF SER SYSTEM USING DSCNN ARCHITECTURE WITH 1500 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	83.3	0	15	1.7
	Sad	1	95	1.4	2.6
	Neutral	0	0	96.6	3.4
	Happy	2.6	18	2.8	76.6

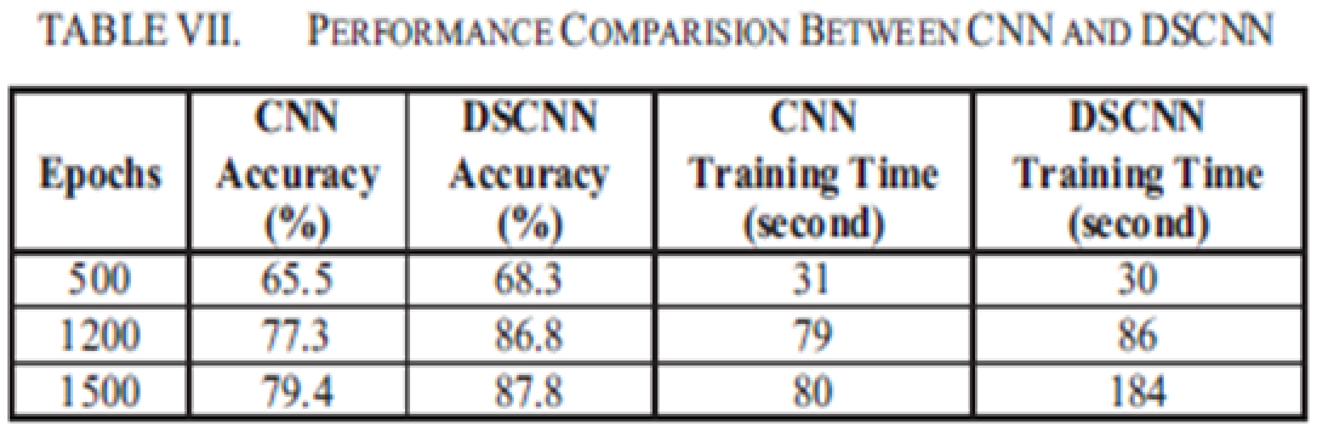


TABLE VII. PERFORMANCE COMPARISION BETWEEN CNN AND DSCNN

Epochs	CNN Accuracy (%)	DSCNN Accuracy (%)	CNN Training Time (second)	DSCNN Training Time (second)
500	65.5	68.3	31	30
1200	77.3	86.8	79	86
1500	79.4	87.8	80	184

- The focus of Speech Emotion Recognition research is to design proficient and robust methods to recognize emotions. In this paper, we have modified the recently proposed algorithm.
- Deep Stride Convolutional Neural Networks (DSCNN) by decreasing the number of convolutional layers with different sizes.
- This network completely excludes the pooling layers instead makes use of special strides for decreasing the dimensionality of feature maps. Two experiments were carried out to check the effectiveness of the state-of-art model of CNN and DSCNN.
- The input to the models was spectrograms generated from the speech database. 87.8% of accuracy was obtained for DSCNN and 79.4% for CNN.

Literature Survey- Rahul Roshan G

Paper - 1 : "Contextual Emotion Detection in Text using Deep Learning and Big Data"

Introduction:

- The paper discusses the importance of emotional detection in dialogue systems and how this can be achieved through the use of big data and deep learning algorithms.
- The paper describes the need for machines to comprehend human behavior and respond emotionally to users in a human-like manner.
- The paper discusses the use of the LSTM model to detect emotions, and how emoticons can be used to express emotions in text messages.

Proposed model:

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.

Literature Survey- Rahul Roshan G

Pros:

- 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.

Cons:

- It is important to note that this pre-processing 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 doesnt deal with emojis and the emotion the paper detected is very less (happy, sad, angry and others)

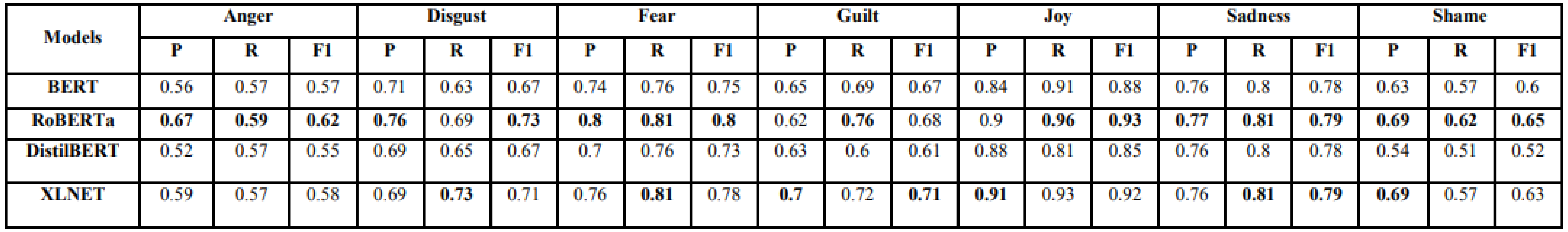
Result

- The paper normalized this data through various techniques and corrected it in data spelling correction.
- Among the various Models, emotion detected reaches the highest accuracy of 0.85 in our papers.

Literature Survey- Rahul Roshan G

COMPARATIVE ANALYSES OF BERT, ROBERTA, DISTILBERT, AND XLNET FOR TEXT-BASED EMOTION RECOGNITION

Table 2 Comparison of Precision, Recall And F1-scores of BERT, RoBERTa, DistilBERT, and XLNET on the ISEAR Dataset



Models	Anger			Disgust			Fear			Guilt			Joy			Sadness			Shame		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT	0.56	0.57	0.57	0.71	0.63	0.67	0.74	0.76	0.75	0.65	0.69	0.67	0.84	0.91	0.88	0.76	0.8	0.78	0.63	0.57	0.6
RoBERTa	**0.67**	**0.59**	**0.62**	**0.76**	0.69	**0.73**	**0.8**	**0.81**	**0.8**	0.62	**0.76**	0.68	0.9	**0.96**	**0.93**	**0.77**	**0.81**	**0.79**	**0.69**	**0.62**	**0.65**
DistilBERT	0.52	0.57	0.55	0.69	0.65	0.67	0.7	0.76	0.73	0.63	0.6	0.61	0.88	0.81	0.85	0.76	0.8	0.78	0.54	0.51	0.52
XLNET	0.59	0.57	0.58	0.69	**0.73**	0.71	0.76	**0.81**	0.78	**0.7**	0.72	**0.71**	**0.91**	0.93	0.92	0.76	**0.81**	**0.79**	**0.69**	0.57	0.63

- The implemented models are fine-tuned on the ISEAR data to distinguish emotions into anger, disgust, sadness, fear, joy, shame, and guilt.

- Using the same hyperparameters, the recorded model accuracies in decreasing order are 0.7431, 0.7299, 0.7009, 0.6693 for RoBERTa, XLNet, BERT, and DistilBERT, respectively.

- BERT is the original transformer-based language model, RoBERTa is an improved version of BERT with a larger and more diverse training corpus, and DistilBERT is a distilled version of BERT that is faster and more memory-efficient. All three models can be fine-tuned on a range of NLP tasks and have been shown to achieve state-of-the-art performance on many benchmarks. XLNet is based on the Transformer architecture, which uses self-attention mechanisms to capture the relationships between words in a sentence.

Literature Survey - Sohan M H

Paper : "Emotionally charged text classification with deep learning and sentiment semantic"

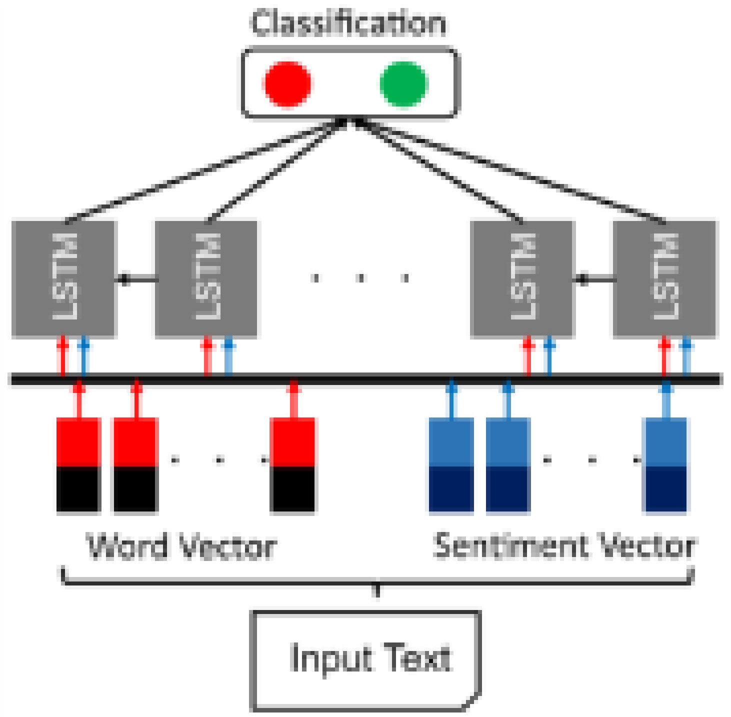
Introduction:

- "Emotionally charged text classification with deep learning and sentiment semantic" presents a method for classifying emotionally charged text using deep learning and sentiment semantics.
- The paper incorporates semantic information from pre-trained GloVe vectors , and sentiment information from SentiWordNet (lexicons) to mirror the way human comprehends text.

Model used:

- The proposed method uses a text classifier that uses a dual-modality of information extraction and a long short-term memory recurrent neural network (LSTM) for the classification. Firstly, a word embedding feature is extracted from the pre-trained model. Next, the emotion of the text is extracted from the sentiment network.
- Finally, the features are combined to classify the text. An LSTM is a type of artificial neural network with self-connection and nodes mad up of gated memory blocks.

Literature Survey - Sohan M H



Pros :

- The proposed method achieves state-of-the-art performance on emotionally charged text classification.
- The use of sentiment semantics (i.e., SentiWordNet lexicon) improves the accuracy of emotion classification.

Cons :

- The proposed method only focuses on emotionally charged text classification, and it may not generalize well to other types of text classification tasks.
- The model doesn’t deal with pictorial data (like emojis,etc).
- The method requires a large amount of labeled data for training, which may be difficult to obtain in some domains.

Literature Survey - Sohan M H

Sentiment Information:

- WordNet is a type of database that categorizes words into groups that mean the same thing. By looking at how these groups are connected and how many words are in each group, we can figure out how similar different words are.
- SentiWordNet (a lexicon) is an extension of WordNet that takes things a step further by giving each group a score that tells us how positive, negative, or neutral the words in that group are.

Summary:

- The authors evaluated the proposed method on two datasets: SemEval-2017 Task 4 and EmoReact. The results showed that the proposed method outperformed several state-of-the-art methods for both datasets.
- Specifically, the proposed method achieved an F1 score of 0.659 on SemEval-2017 Task 4 and an F1 score of 0.718 on EmoReact, which are significantly higher than the best-performing baseline methods.
- In conclusion, the proposed method in the paper "Emotionally charged text classification with deep learning and sentiment semantic" presents a promising approach to emotionally charged text classification using deep learning and sentiment semantics .

Design Approach

Iterative Design Approach

1. Gather requirements: Understand the user requirements and project goals.
2. Design: Create a design for the system architecture, including the emotion detection and speech conversion models.
3. Implement: Implement the models and integrate them into the system architecture.
4. Test: Test the system using sample inputs and evaluate the accuracy of emotion detection and speech conversion.
5. Evaluate: Analyze the results of testing and gather feedback from users to identify areas of improvement.

Benefits:

- Allows for continuous improvement and refinement of the system.
- Can help identify and address issues early in the development process.

Drawbacks:

- Can be time-consuming and costly.
- May require significant changes to the system architecture if major issues are identified.

Design Approach

Agile Design Approach

1. Plan: Create a high-level plan for the system architecture, including the emotion detection and speech conversion models.
2. Develop: Develop the models and integrate them into the system architecture in small, incremental stages.
3. Test: Test the system using sample inputs and gather feedback to evaluate the system's effectiveness and usability.
4. Evaluate: Analyze the results of testing and use them to guide further development.

Benefits:

- Allows for flexibility and adaptability in response to changing user needs or system requirements.
- Enables the development team to focus on high-priority features and functionality.

Drawbacks:

- Can lead to a lack of clarity around the overall system architecture and goals.
- May result in technical debt if short-term solutions are prioritized over long-term considerations.

Design Constraints, Assumptions & Dependencies

Design constraints and assumptions:

1. Availability of labeled speech and text dataset.

2. Processing power and resources.

3. Accuracy of emotion recognition.

4. English Language with a UK/US accent.

5. Five proposed emotions (happy, sad, neutral, angry, sarcastic).

6. Annotated form of text (coversational format of text, role based dialogue format) .

Design Constraints, Assumptions & Dependencies

Dependencies:

1. Availability of labeled speech and text dataset (accuracy).

2. Accuracy of the emotion recognition model (user adoption).

3. Availability of processing power and resources (project delays and quality issues).

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.

Design Details

- Novelty:
  - The proposed system recognizes sarcasm as an emotion and uses new machine learning models to accurately convey it in generated speech.
  - It is a departure from traditional emotion recognition systems that do not recognize sarcasm as a distinct emotion.

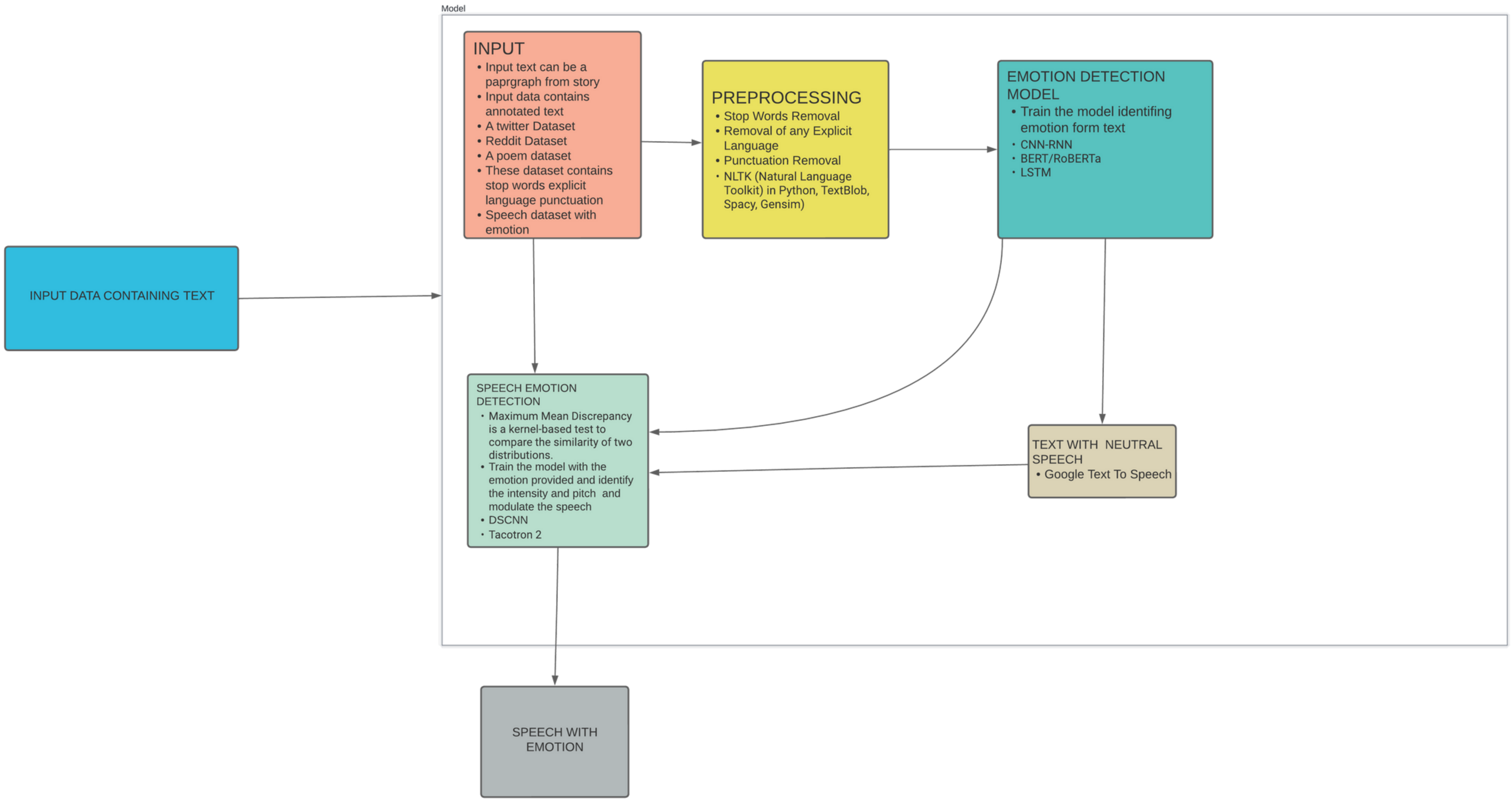
- Performance:
  - To ensure that the system performs optimally, we will need to evaluate its performance under various conditions, such as different types of input text, accents, and languages.
  - We will aim to optimize the system for speed, accuracy, and efficiency to ensure that it can generate emotional speech in real-time.

- Reliability:
  - The system will need to be reliable to ensure that it can operate effectively and consistently under different conditions.
  - We will aim to reduce the system's susceptibility to errors, bugs, and crashes and provide robust error handling and recovery mechanisms.

Proposed Methodology / Approach

- The input data consist of dataset from various sources like social media, poetry, paragraph .Preprocessing is done on the text dataset like removing stop words, explicit words,punctuation.
- The labeled text data being provided as input to the emotion detection training model. The model analyzes the text and identifies the corresponding emotion. This output is then passed on to the speech emotion detection model.
- The text is converted to neutral speech using google tts . Both the neutral speech and emotion from the emotion detection model is passed to the sppech emotion detection model.
- The speech emotion detection model adjusts the speech to reflect the appropriate emotion based on the input text and the corresponding emotion identified by the emotion detection training model.
- The resulting output is speech that accurately reflects the emotional content of the input text.

Architecture



INPUT DATA CONTAINING TEXT

INPUT
- Input text can be a paprgraph from story
- Input data contains annotated text
- A twitter Dataset
- Reddit Dataset
- A poem dataset
- These dataset contains stop words explicit language punctuation
- Speech dataset with emotion

PREPROCESSING
- Stop Words Removal
- Removal of any Explicit Language
- Punctuation Removal
- NLTK (Natural Language Toolkit) in Python, TextBlob, Spacy, Gensim)

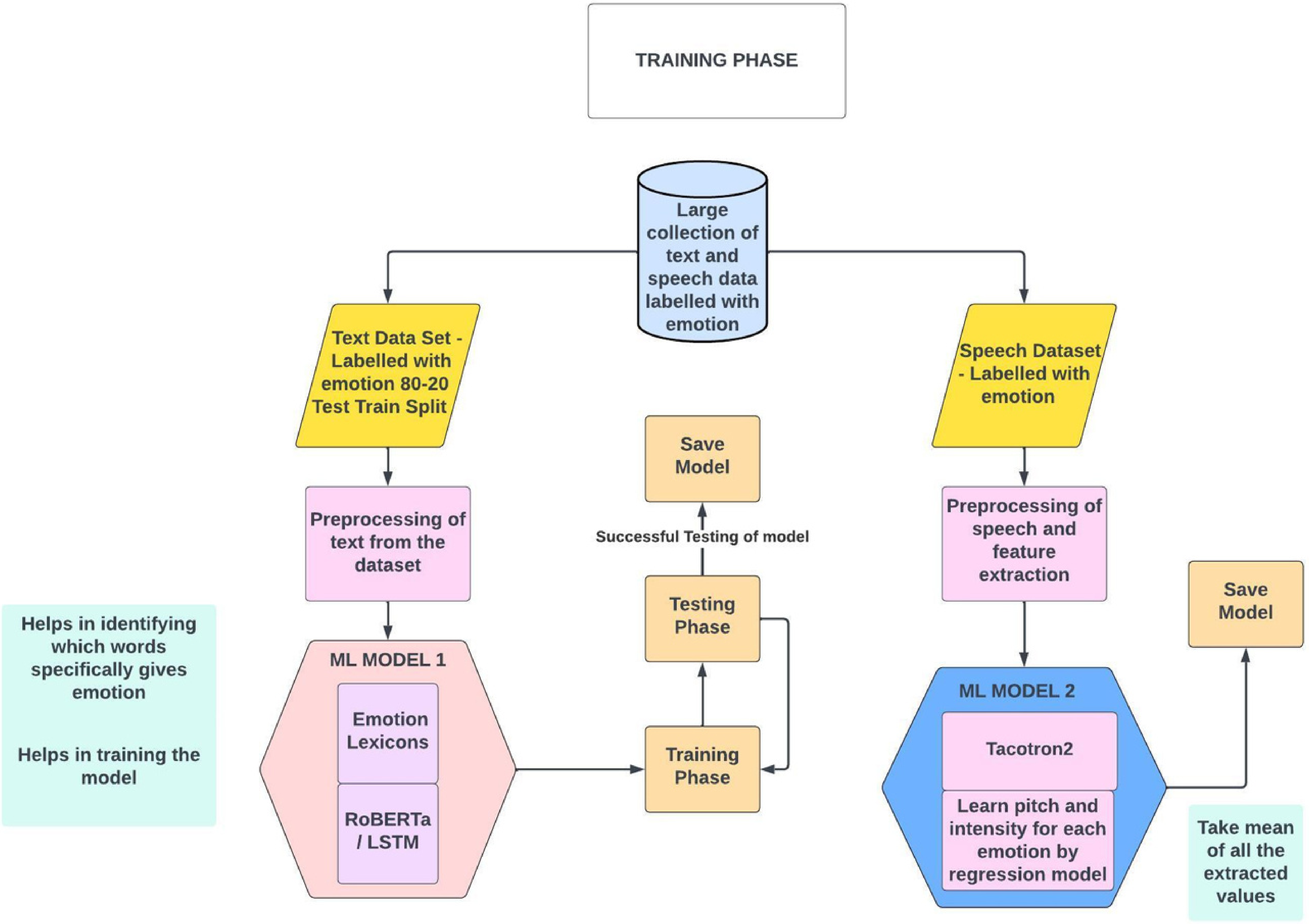
EMOTION DETECTION MODEL
- Train the model identifing emotion form text
- CNN-RNN
- BERT/RoBERTa
- LSTM

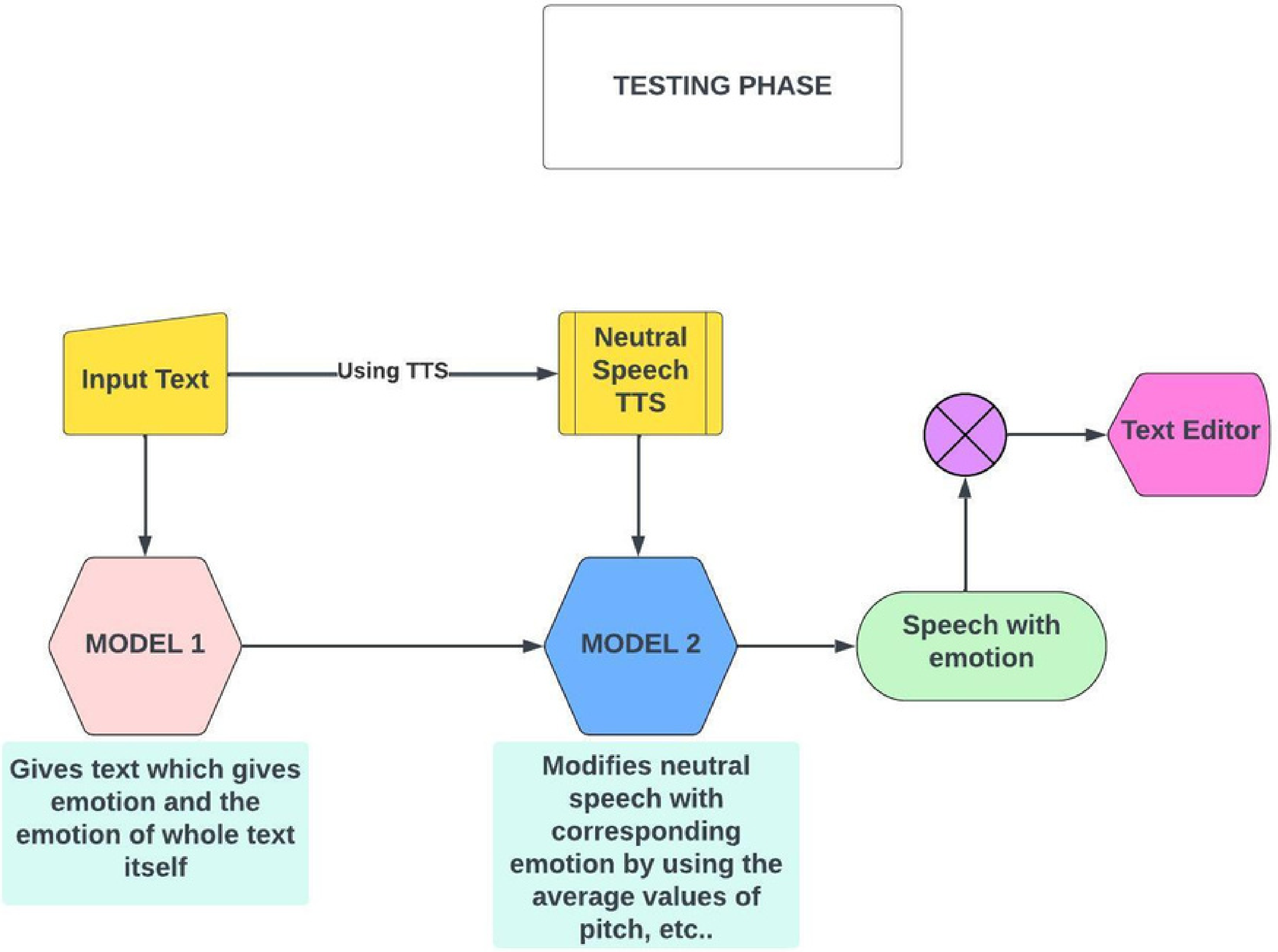
SPEECH EMOTION DETECTION
- Maximum Mean Discrepancy is a kernel-based test to compare the similarity of two distributions.
- Train the model with the emotion provided and identify the intensity and pitch and modulate the speech
- DSCNN
- Tacotron 2

TEXT WITH NEUTRAL SPEECH
- Google Text To Speech

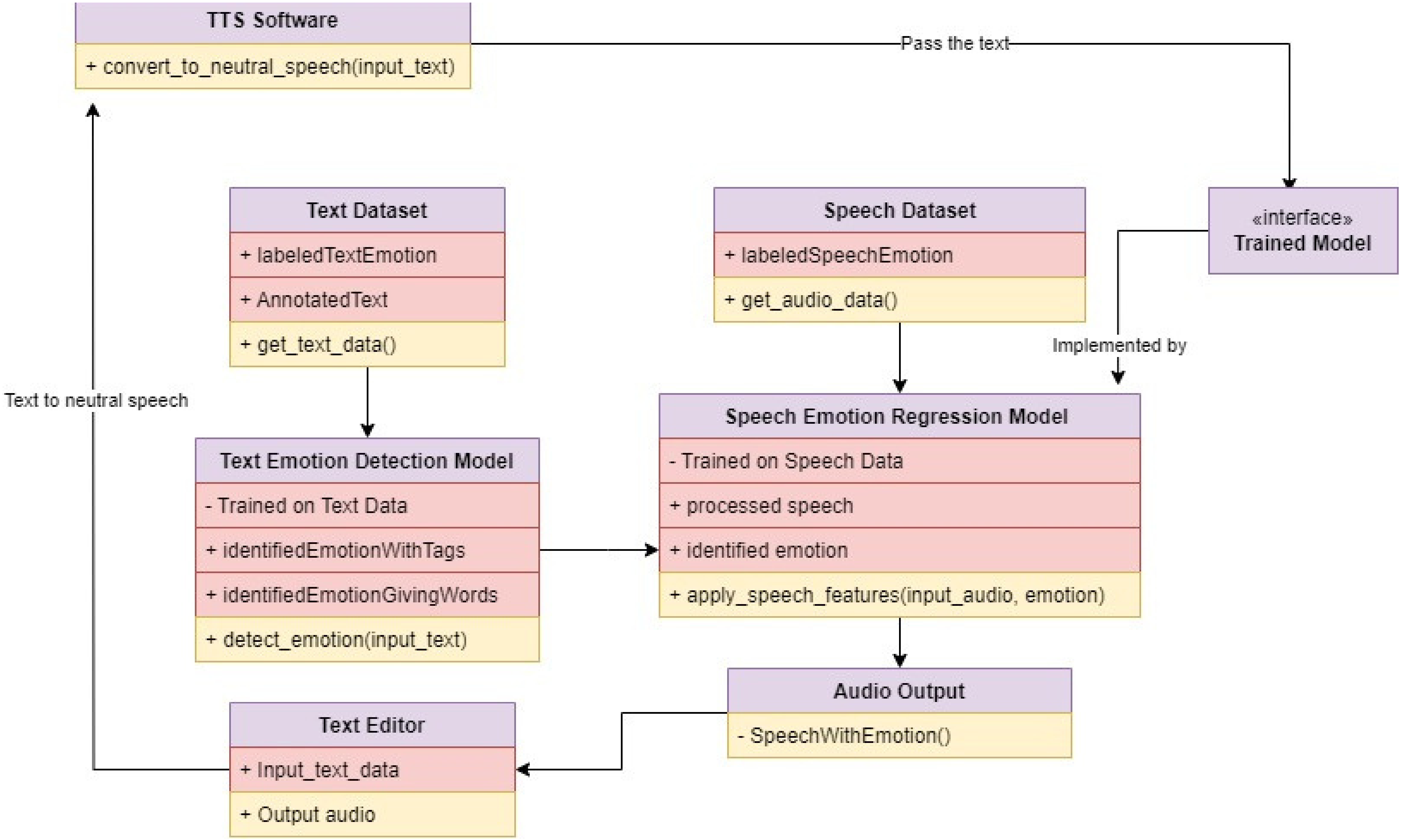
SPEECH WITH EMOTION

Sequence


TRAINING PHASE
Large collection of text and speech data labelled with emotion
Text Data Set - Labelled with emotion 80-20 Test Train Split
Speech Dataset - Labelled with emotion
Save Model
Successful Testing of model
Preprocessing of text from the dataset
Preprocessing of speech and feature extraction
Testing Phase
Save Model
Helps in identifying which words specifically gives emotion
Helps in training the model
ML MODEL 1
Emotion Lexicons
RoBERTa / LSTM
Training Phase
ML MODEL 2
Tacotron2
Learn pitch and intensity for each emotion by regression model
Take mean of all the extracted values


TESTING PHASE
Input Text
Using TTS
Neutral Speech TTS
Text Editor
MODEL 1
MODEL 2
Speech with emotion
Gives text which gives emotion and the emotion of whole text itself
Modifies neutral speech with corresponding emotion by using the average values of pitch, etc..

Master class diagram



TTS Software
+ convert_to_neutral_speech(input_text)

Pass the text

«interface»
Trained Model

Implemented by

Text Dataset
+ labeledTextEmotion
+ AnnotatedText
+ get_text_data()

Speech Dataset
+ labeledSpeechEmotion
+ get_audio_data()

Text to neutral speech

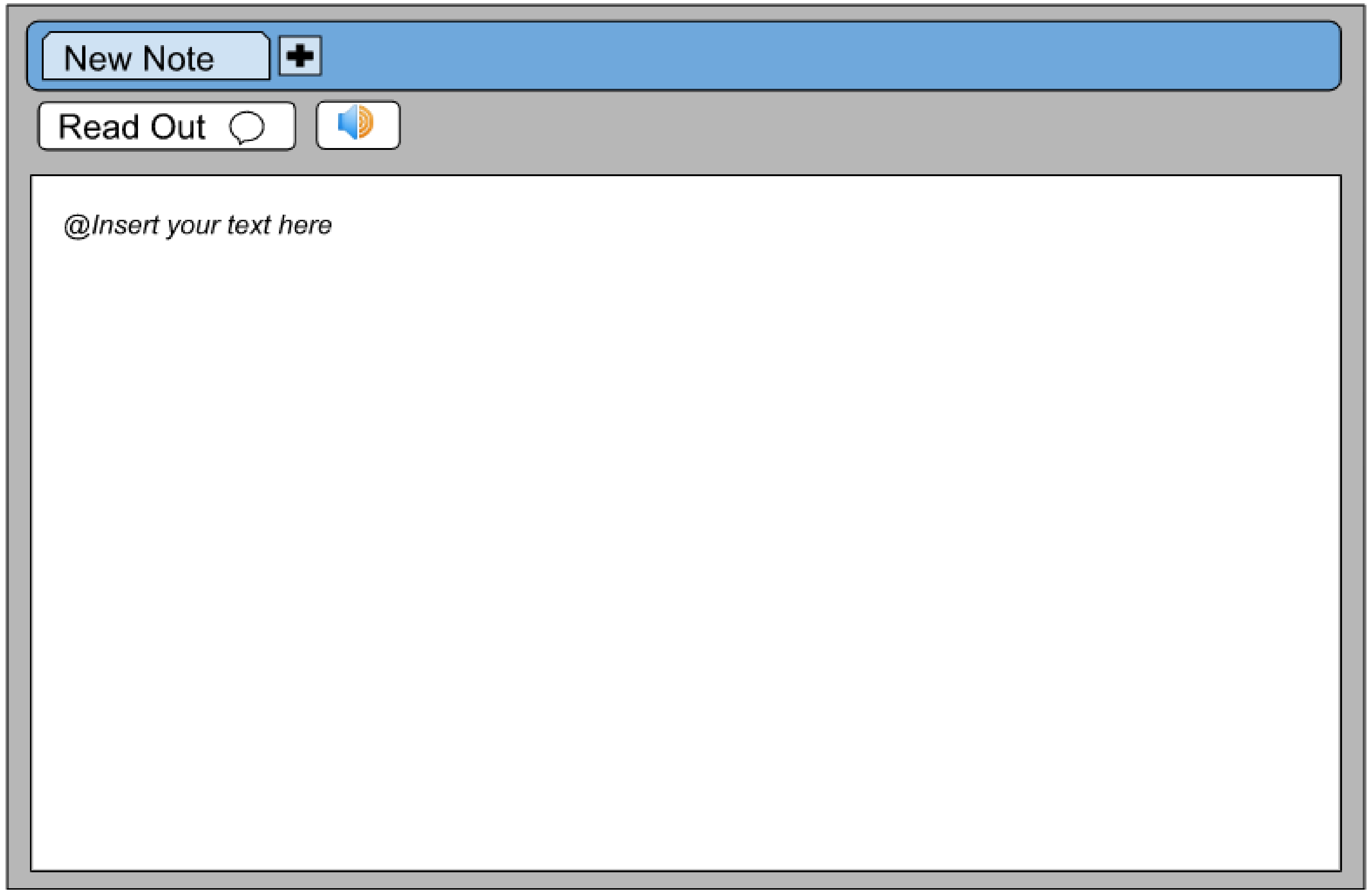
Text Emotion Detection Model
- Trained on Text Data
+ identifiedEmotionWithTags
+ identifiedEmotionGivingWords
+ detect_emotion(input_text)

Speech Emotion Regression Model
- Trained on Speech Data
+ processed speech
+ identified emotion
+ apply_speech_features(input_audio, emotion)

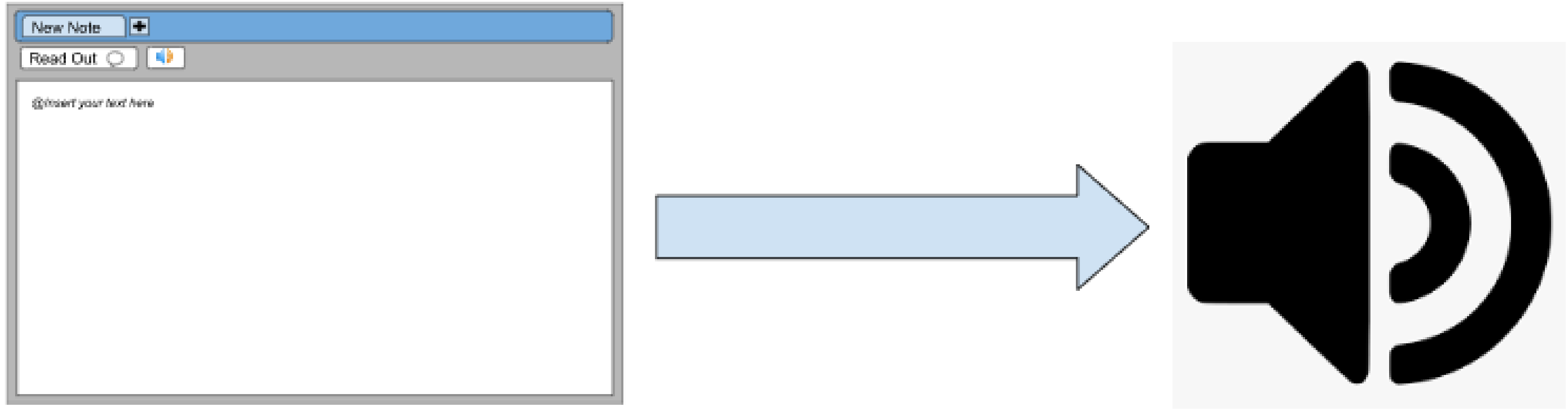
Audio Output
- SpeechWithEmotion()

Text Editor
+ Input_text_data
+ Output audio

User Interface Diagram



External Interfaces



Text Input From the user

Speech is generated with emotion

Technologies Used

- Recurrent neural network (RNN) and Long Short-Term Memory (LSTM) architectures, which are commonly used in TTS systems to model the sequence of input text and generate corresponding speech outputs.
- Google Text-to-Speech (TTS), a cloud-based API that can be used to generate speech from text inputs in a variety of languages and accents.
- Tacotron, a neural network-based TTS system that uses an encoder-decoder architecture with attention mechanisms to generate high-quality, natural-sounding speech from text inputs.
- Speech emotion recognition (SER) technology, such as a CNN model trained on a labeled dataset of emotional speech, which can be used to classify the emotional content of input speech samples and adjust the emotional tone of the synthesized speech.
- Natural language processing (NLP) tools, such as tokenization and part-of-speech tagging, which can be used to preprocess input text and extract relevant features for input to the TTS model.
- Machine learning (ML) technology, such as training the TTS and SER models on large datasets of speech and emotional labels, to improve the accuracy and performance of the models.
- Signal processing technology, such as spectral analysis and filtering, which can be used to process the speech signals and modify the emotional content of the generated speech.
- Overall, the combination of these technologies can be used to create a speech synthesis system that can generate speech with controllable emotional content, allowing users to adjust the emotional tone of the synthesized speech to suit their needs.

Project Progress

What is the project progress so far?

- In this project, we are now ready with all the architecture, high-level design, algorithms and tools required, methodologies from literature survey, etc...
- We are aware of the limitations and expected deliverables for the project.
- We have also looked for the datasets, both text and speech labelled with emotion which are annotated. We are ready with tools and other requirements and particularly the flow of the project.
- Now, we are to just filter our datasets used for training and then start with the implementation.
- We also have a text editor ready-in-hand for the model to be integrated with.

What is the percentage completion of the project?

- Therefore referring to the above progress we feel that we have completed around 30 % of the project

Summary of work done in Capstone Phase-1

The project has completed the necessary preparation stages, including the development of the architecture, algorithms, and tools required, as well as a literature survey and methodology review.

The team has also identified and obtained the necessary datasets labeled with emotion. The next step is to filter the datasets and begin implementing the project.

A text editor has been selected for integration with the model.

Project Plan for Capstone Phase-2

- Sarcasm detection and Speech:The team is currently working on incorporating sarcasm into the model, and has obtained appropriate emotion-labelled datasets which are being filtered before implementation.
- Data Preprocessing: The team will need to preprocess the datasets by cleaning and filtering the data, converting the text to the appropriate format for input to the TTS model, and preparing the speech data for use in training the emotion recognition model.
- Model Training: The team will train the TTS model and the emotion recognition model on the preprocessed datasets using appropriate machine learning algorithms and techniques. This may involve experimenting with different architectures and hyperparameters to optimize the performance of the models.
- Evaluation and Testing: The team will need to evaluate the performance of the system by testing it on a range of input text and speech samples and measuring the accuracy of the emotion recognition and speech synthesis components. The team may also conduct user studies to gather feedback on the usability and effectiveness of the system.
- Refinement and Optimization: Based on the evaluation results, the team will refine and optimize the models and algorithms to improve the performance of the system. This may involve further data preprocessing, model training, and testing iterations.
- Documentation and Reporting: The team will document the project implementation process, including the design decisions, implementation details, and evaluation results, in a final project report. The team may also create user manuals or technical documentation to support the deployment and use of the system.

References

]1[Cai, Xiong, et al. "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). IEEE, 2021.

[2] 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

[3] 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.

[4] 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) ISBN Information:INSPEC Accession Number: 20133021
DOI: 10.1109/ICWT50448.2020.9243622

Thank You