UE20CS390A – Capstone Project Approval

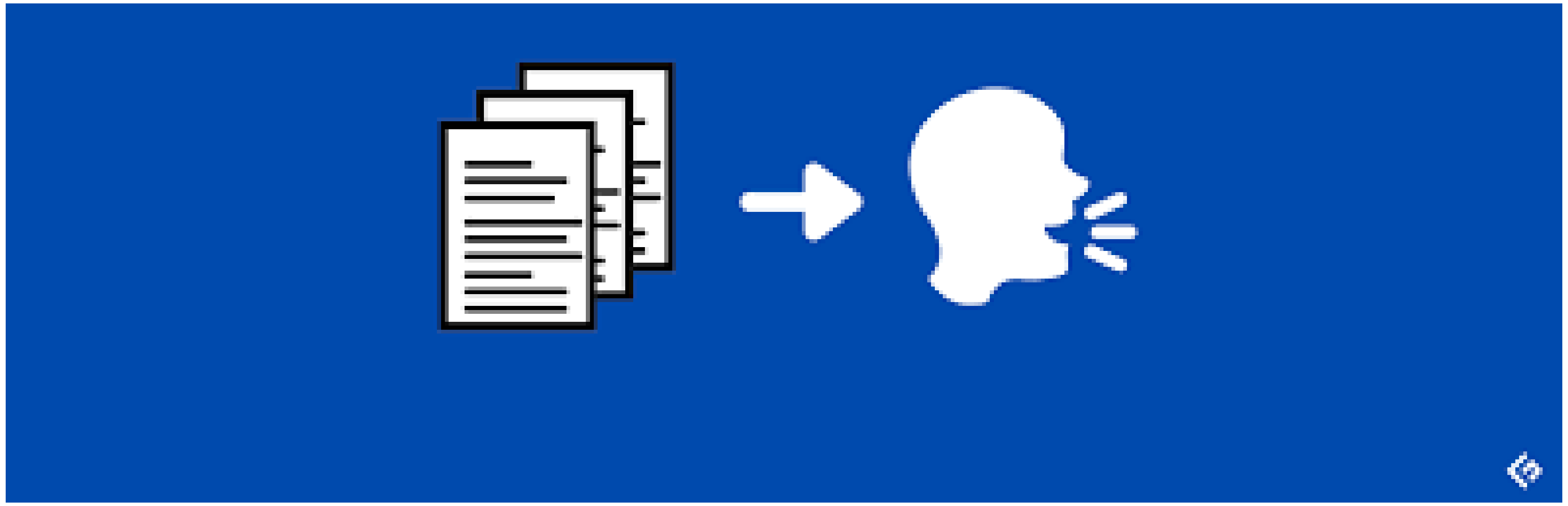
Project Title: FeelSpeak: Generating Emotional Speech with Deep Learning
Project ID: PW23_VRB_07
Project Guide: V R Badri Prasad
Project Team: 235_320_355_362

Outline

- Problem Statement
- Scope and Feasibility study
- Applications/Use cases
- Expected Deliverables
- Capstone (Phase-I & Phase-II) Project Timeline
- Any other information

Problem Statement

FeelSpeak: Generating Emotional Speech with Deep Learning



Scope and Feasibility study

INTRODUCTION :

Speech synthesis technology has made significant advances in recent years, and it has the potential to greatly impact various industries and applications, including virtual assistants, voice-controlled devices, and accessibility technology. One important aspect of speech synthesis is the generation of speech with emotions, which is necessary to enhance the naturalness and expressiveness of the speech. In this study, we will examine the feasibility of generating speech with emotions from text using machine learning techniques.

SCOPE :

The scope of this feasibility study is to evaluate the potential of using machine learning to generate speech with emotions from the text. This includes identifying the data and models required, evaluating the performance of different approaches, and determining the challenges and limitations of the technology.

Scope and Feasibility study

REQUIREMENTS :

To carry out this feasibility study, the following requirements are necessary:

- A large, diverse, and high-quality dataset for training the model, including text and speech with corresponding emotions with our modifications (Challenging Task)
- Access to deep learning models and tools for training the model and evaluating its performance
- A user interface for inputting text and listening to the generated speech with emotions
- Evaluation metrics to assess the performance of the model, such as accuracy, naturalness, and expressiveness

EVALUATIONS :

The evaluation of the feasibility of generating speech with emotions from text using machine learning will involve several steps, including:

- Preprocessing and cleaning the training data to ensure its quality and consistency
- Training the model using the preprocessed data and evaluating its performance using the evaluation metrics
- Comparing the performance of the model with existing solutions and identifying areas for improvement
- Carrying out user studies to get feedback on the model's performance and usability from real users

Scope and Feasibility study

EXTRACTIONS :

Based on the results of the feasibility study, we will draw conclusions on the potential of using machine learning to generate speech with emotions from text, including:

- Whether the technology is capable of generating speech with emotions from text with acceptable accuracy and naturalness
- Whether the technology is able to overcome the challenges and limitations of existing solutions
- Recommendations for future work to improve the technology and its applications

In conclusion, this feasibility study will provide valuable insights into the potential of using machine learning
to generate speech with emotions from text, and it will serve as a basis for further development and deployment of the technology.

Applications/Use cases

Applications:

Virtual assistants: Virtual assistants that use speech as a primary mode of communication can benefit from being able to convey emotions, such as friendliness, empathy, and urgency, to enhance the user experience.

Customer service: Customer service representatives often use phone or chat systems to communicate with customers. Generating speech with emotions can help these representatives to better convey their tone and empathy towards customers, improving the overall customer experience.

Assistive technology for people with disabilities: People with disabilities, such as those with speech or hearing impairments, can benefit from assistive technology that generates speech with emotions. This can help to improve communication and interaction with others.

Entertainment: The technology can be used in the entertainment industry to generate speech for virtual characters or to create more engaging audio experiences for users.

Education: The technology can be used in educational settings to create more engaging and expressive speech for language learning apps, educational games, and other interactive educational experiences.

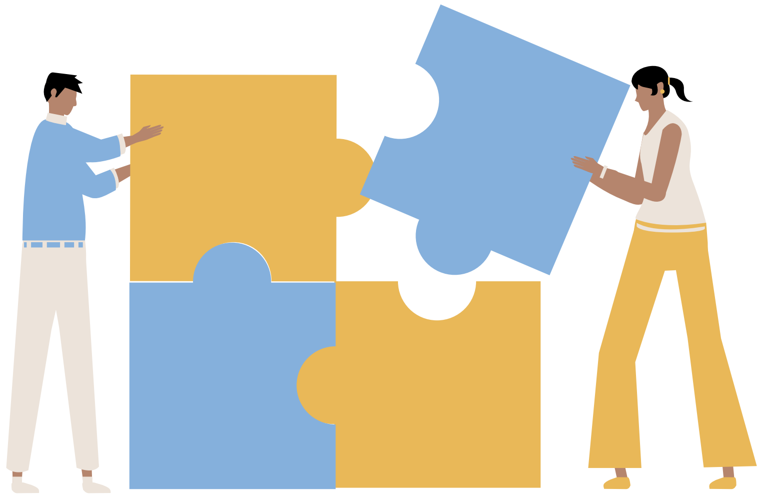
Expected Deliverables

Capstone-I deliverables

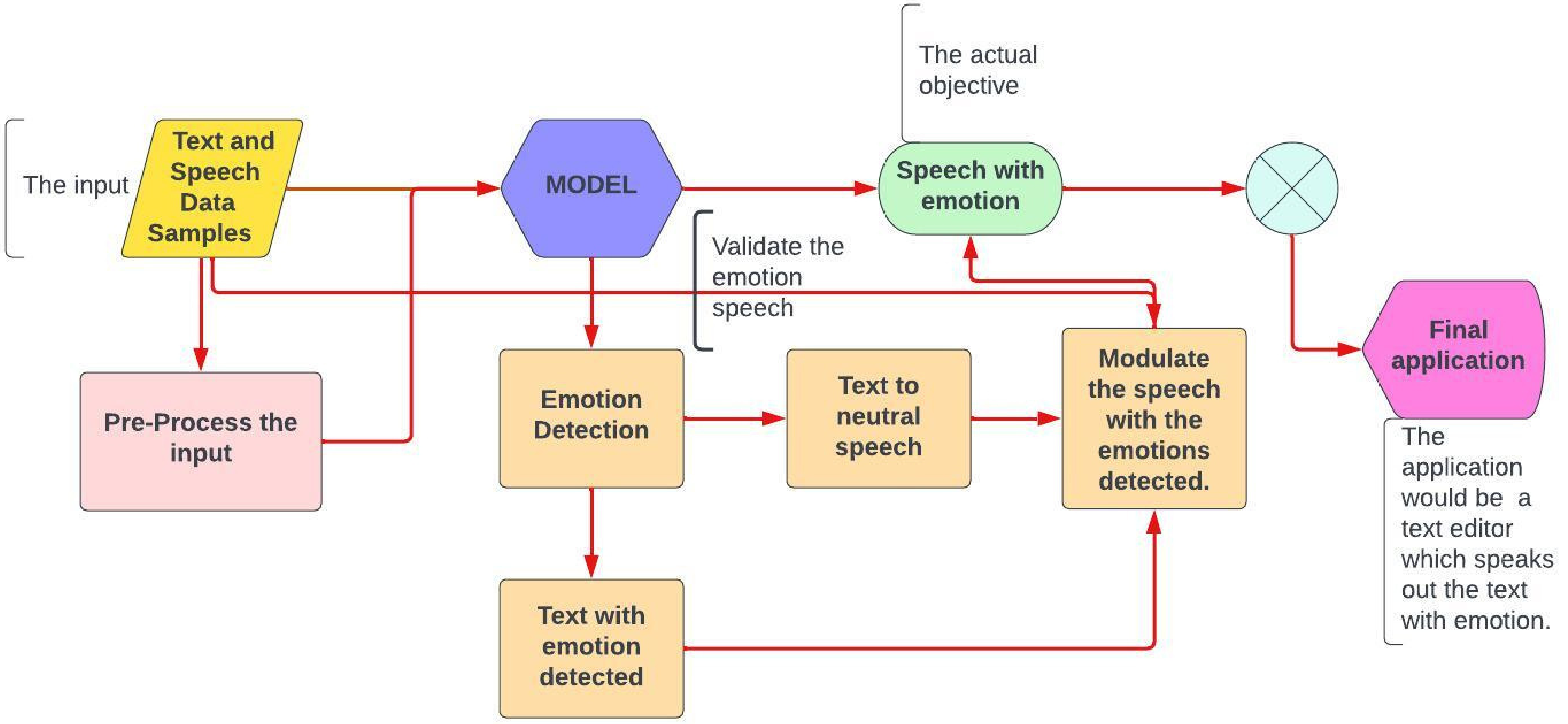
- Exhaustive Literature Study
- Checking for the feasibility of the project
- Its Application in real world
- High Level Design
- Sample Coding and Unit Testing of basic models

Capstone-II deliverables

- Low Level Design
- Preprocessing of input data
- Emotion Detection from text
- Converting text to neutral speech
- Speech Modulation
- Incorporating with text editor
- Creating test cases
- Sarcasm Detection



Flow Diagram - Training


The input
Text and Speech Data Samples
MODEL
The actual objective
Speech with emotion
Validate the emotion speech
Pre-Process the input
Emotion Detection
Text to neutral speech
Modulate the speech with the emotions detected.
Final application
The application would be a text editor which speaks out the text with emotion.
Text with emotion detected

Exhaustive Literature Survey

Paper : Conference

Title : Emotion Controllable Speech Synthesis Using Emotion-Unlabeled Dataset with the Assistance of Cross-Domain Speech Emotion Recognition

Objectives : Cross Domain SER & TTS Model

Summary : The cross-domain SER model is pre-trained on the emotion labeled SER dataset (as the source domain) and the emotion-unlabeled TTS dataset (as the target domain). Then, the soft emotion labels of the TTS dataset are obtained from the softmax output of the trained SER model (the green dashed arrow). Finally, the TTS model and an emotion predictor are jointly trained on the TTS dataset with the emotion labels.

Results :

Year : 2020

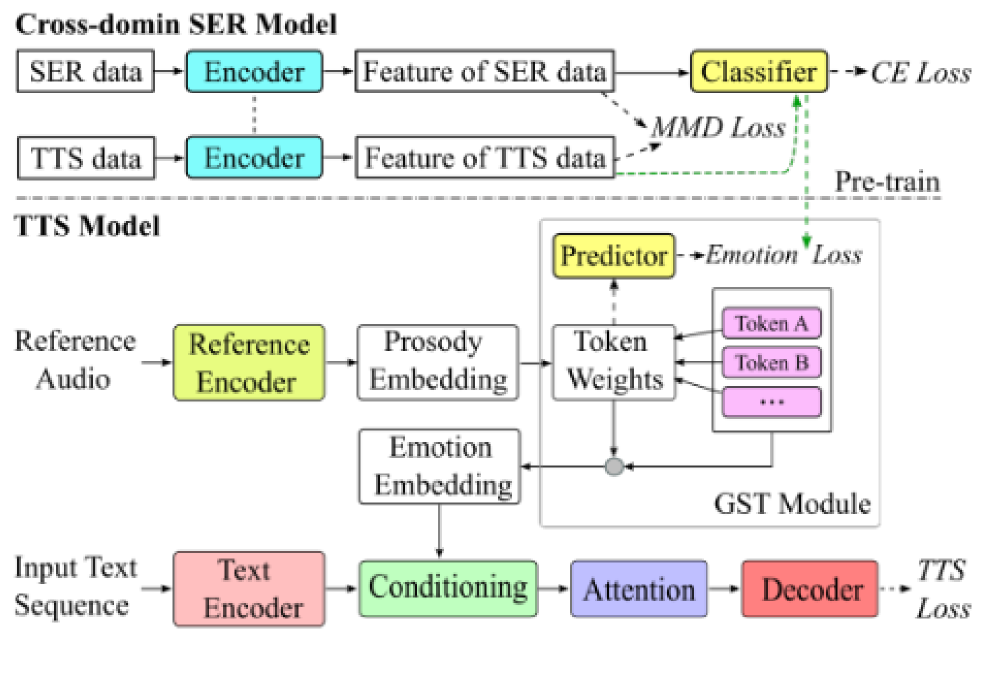


Fig. 1. The overall structure of the cross-domain SER and GST-based TTS model.

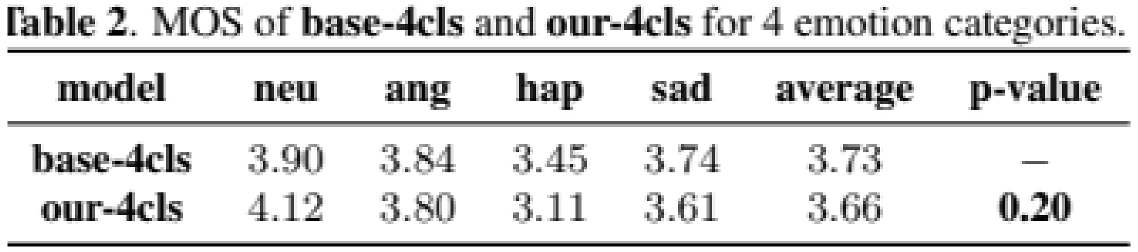


Table 2. MOS of **base-4cls** and **our-4cls** for 4 emotion categories.

model	neu	ang	hap	sad	average	p-value
base-4cls	3.90	3.84	3.45	3.74	3.73	–
our-4cls	4.12	3.80	3.11	3.61	3.66	**0.20**

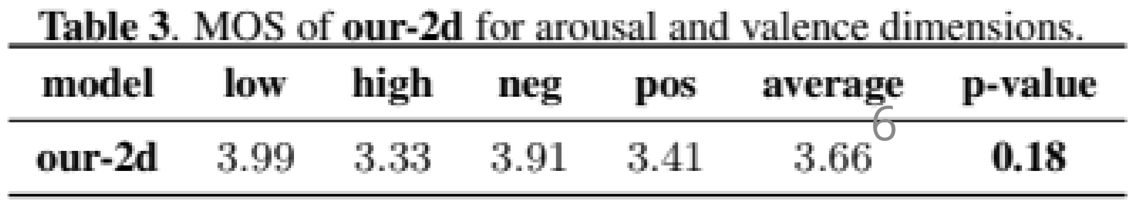


Table 3. MOS of **our-2d** for arousal and valence dimensions.

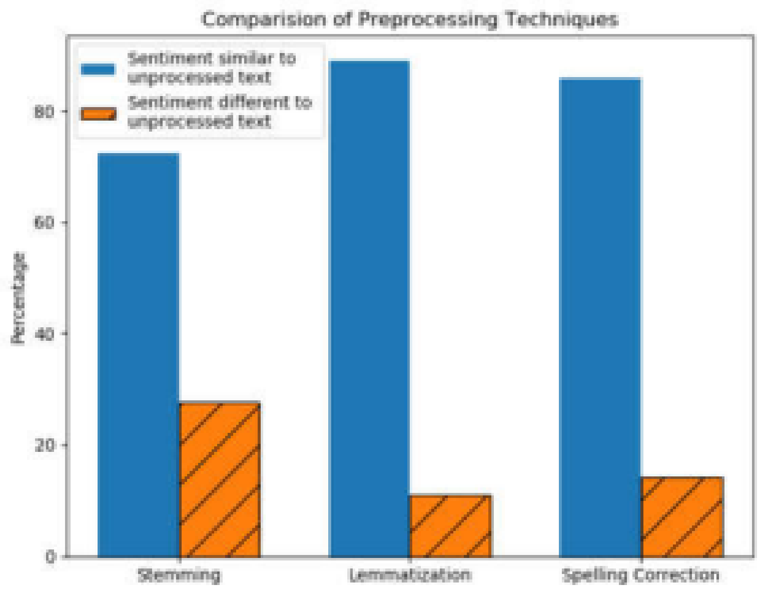
model	low	high	neg	pos	average	p-value
our-2d	3.99	3.33	3.91	3.41	3.66	**0.18**

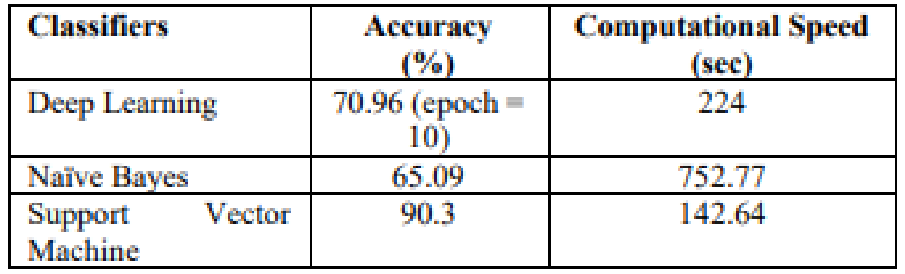
Exhaustive Literature Survey

Paper : 11th IEEE International Conference
Title : Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data
Objectives :

1. Investigate and analyze the most commonly used text data preprocessing techniques for sentiment analysis in social media data.
2. Evaluate the effectiveness of these techniques in improving the sentiment analysis performance.
3. Identify the limitations and challenges of existing text preprocessing techniques.
4. Provide recommendations for the effective preprocessing of text data for sentiment analysis in social media data.

Summary :
The paper "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data" provides a comprehensive analysis of text data preprocessing techniques for sentiment analysis in social media data. The paper investigates and evaluates the most commonly used techniques, identifies their limitations and challenges, and provides recommendations for the effective preprocessing of text data in sentiment analysis of social media data. The paper aims to contribute to the field of sentiment analysis by providing insights into the best practices for text preprocessing and improve the accuracy and reliability of sentiment analysis systems.
Results : As depicted in the images.
Year : 2019





Classifiers	Accuracy (%)	Computational Speed (sec)
Deep Learning	70.96 (epoch = 10)	224
Naïve Bayes	65.09	752.77
Support Vector Machine	90.3	142.64

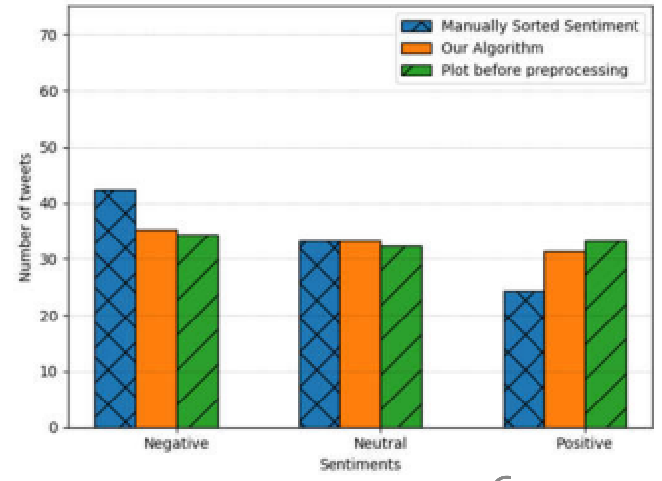
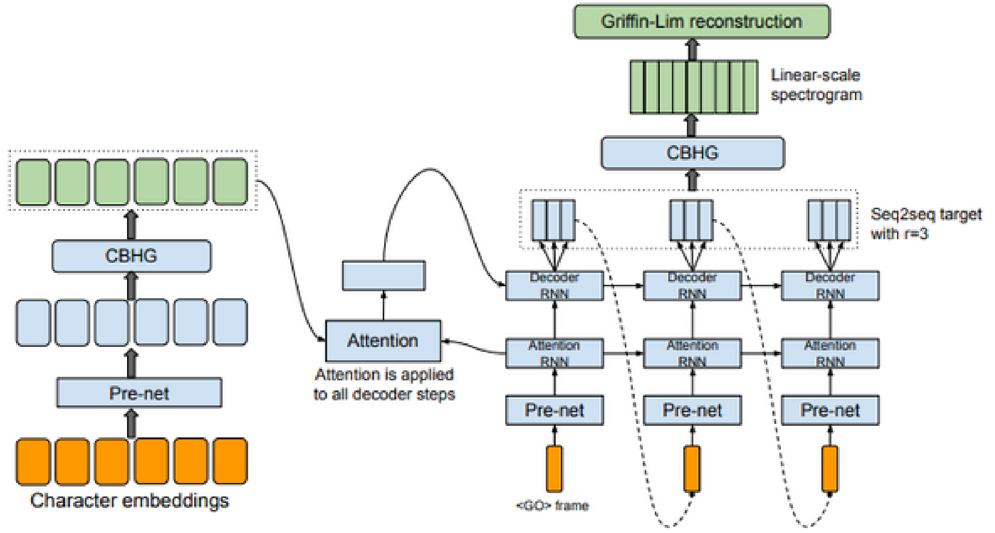


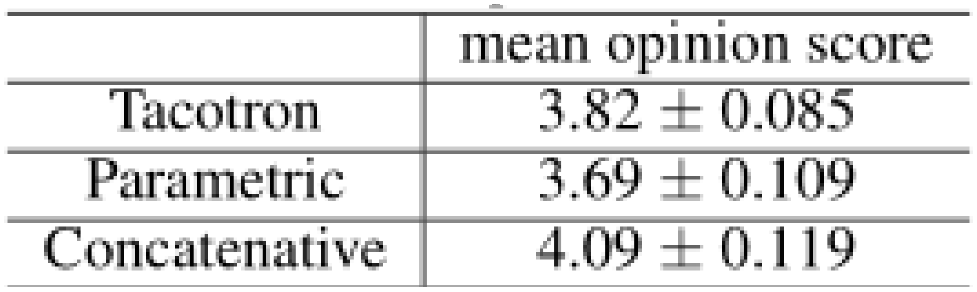
Figure 5: Comparison of Negative, Neutral and Positive Sentiment produced by three different methods

Exhaustive Literature Survey

Paper : Conference of the International Speech Communication Association
Title : Tacotron: Towards End-to-End Speech Synthesis
Objectives: End-End synthesis of speech from text
Summary:

- Tacotron is an deep learning based text-to-speech model that synthesizes speech directly from characters.
- The system consists of two main components: an **encoder-decoder architecture** and a **spectrogram prediction network**. The encoder-decoder architecture processes the input text and generates intermediate representations, which are then passed to the spectrogram prediction network to generate the final speech waveform.
- Tacotron achieves a 3.82 subjective 5-scale mean opinion score on English, outperforming a production parametric system in terms of naturalness.





	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119

Results: MOS Score = 3.82 ± 0.085
Year:2017

Exhaustive Literature Survey

Paper: Second International Conference.
Title: Contextual Emotion Detection in Text using Deep Learning and Big Data.
Objectives: Detect the emotion automatically from our dataset
Summary:

- Detects emotion from textual dialogues and uses of **LSTM** model based for finding out the emotion, detected
- **Supervised** and **monitoring** machine learning methods for emotion detection is a very important part
- **One-third** of the data sources were utilized for **training**, while the other **two-thirds** are being used for **testing**
- **Removal** of out-of-range values, improbable data, missing values invalid words silly or inland data additional spaces
- **Annotated corpus**, like corpus markup, annotation adds value to a corpus.
- **Sword implanting**; a word inserting is an educated portrayal of text where words that have similar importance have a comparable portrayal.
- For **spelling correction** various NLP library.
- **Glove twitter.3B300-implanting** model for **word embedding**
- Has many conversations so **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**.

Results: Accuracy of 0.85
Year: 2022

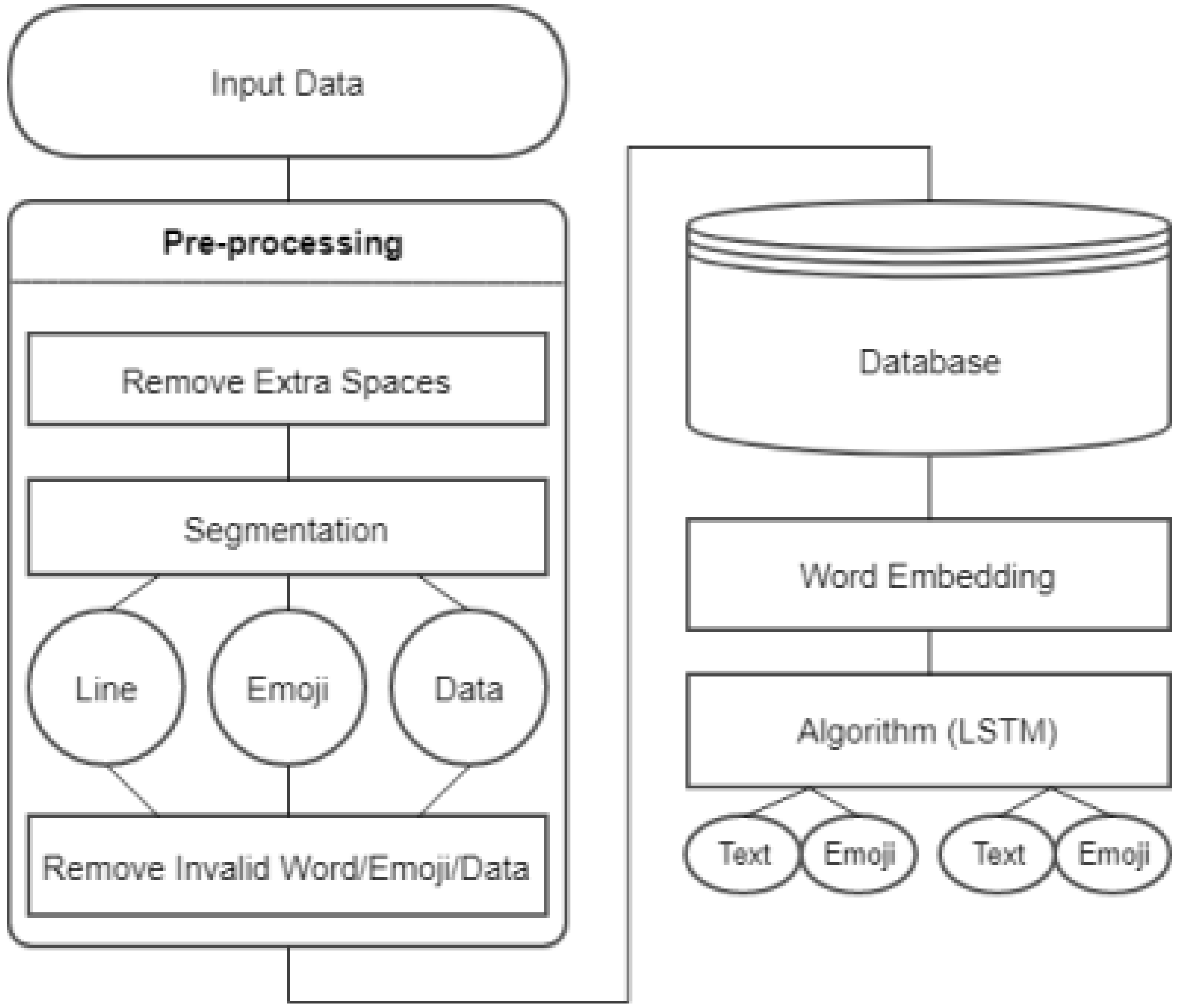
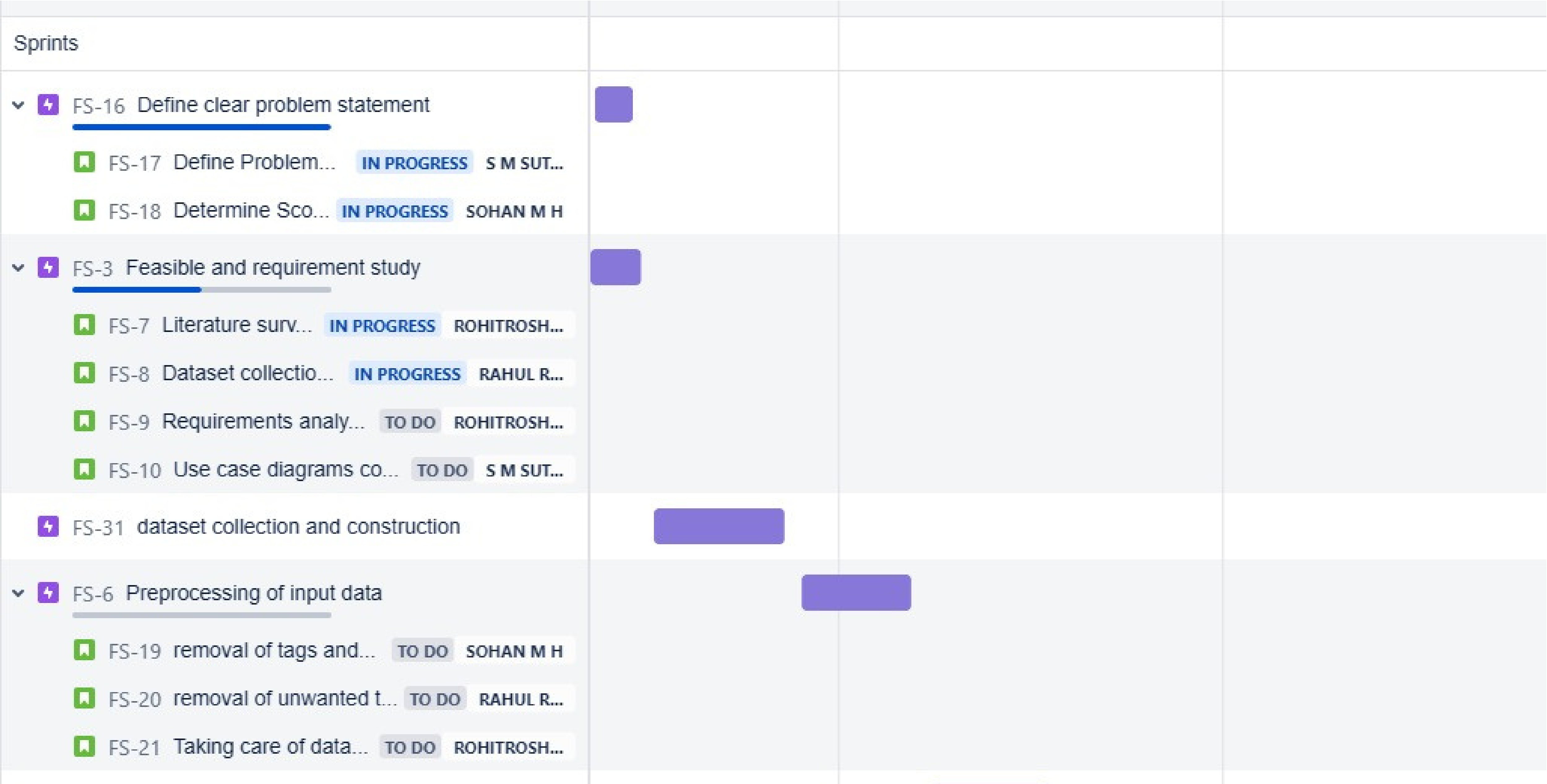


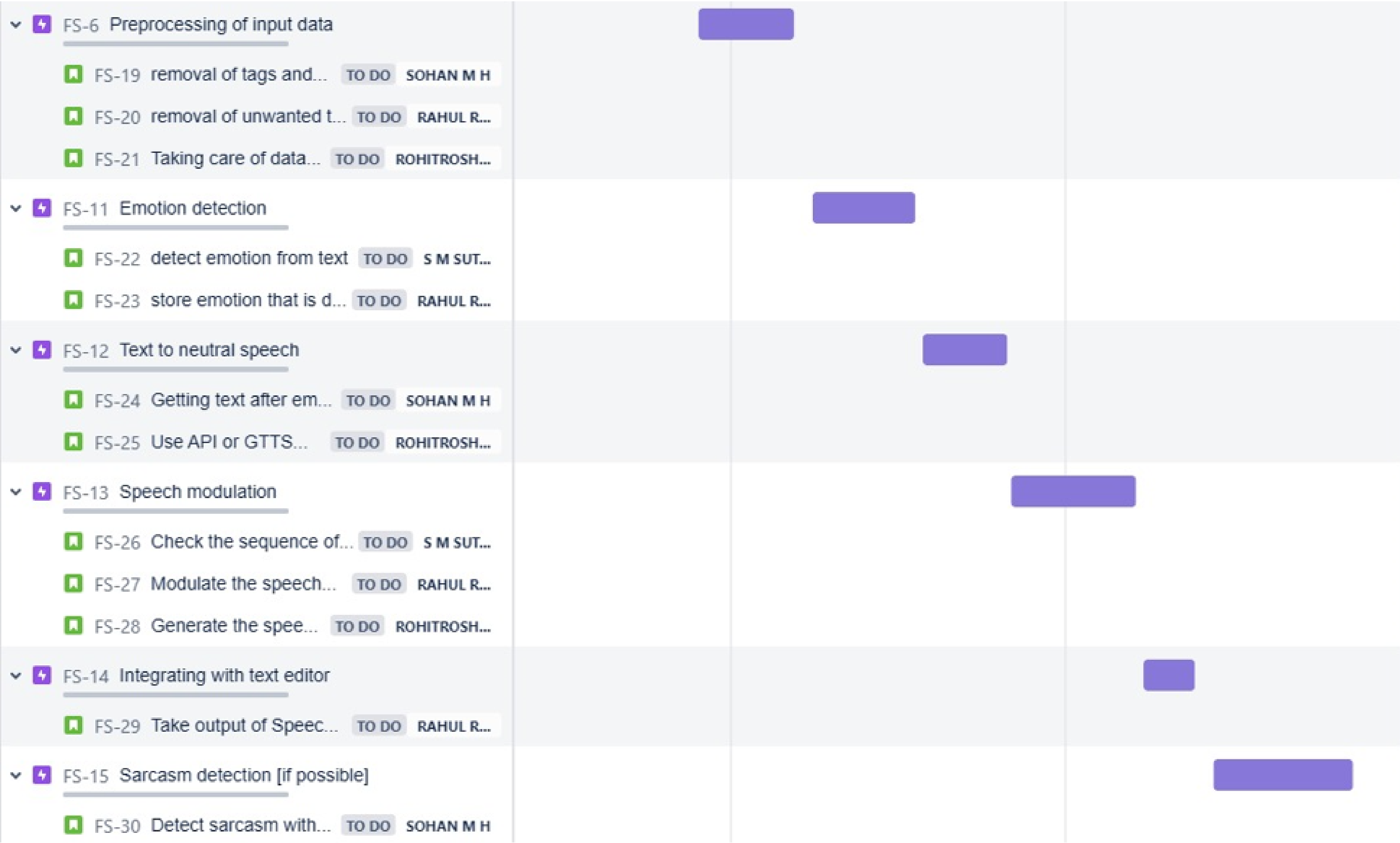
Figure 1. Proposed Model

Capstone (Phase-I & Phase-II) Project Timeline



Sprints		
FS-16 Define clear problem statement		
FS-17 Define Problem...	IN PROGRESS	S M SUT...
FS-18 Determine Sco...	IN PROGRESS	SOHAN M H
FS-3 Feasible and requirement study		
FS-7 Literature surv...	IN PROGRESS	ROHITROSH...
FS-8 Dataset collectio...	IN PROGRESS	RAHUL R...
FS-9 Requirements analy...	TO DO	ROHITROSH...
FS-10 Use case diagrams co...	TO DO	S M SUT...
FS-31 dataset collection and construction		
FS-6 Preprocessing of input data		
FS-19 removal of tags and...	TO DO	SOHAN M H
FS-20 removal of unwanted t...	TO DO	RAHUL R...
FS-21 Taking care of data...	TO DO	ROHITROSH...

Capstone (Phase-I & Phase-II) Project Timeline



FS-6	Preprocessing of input data		
FS-19	removal of tags and...	TO DO	SOHAN M H
FS-20	removal of unwanted t...	TO DO	RAHUL R...
FS-21	Taking care of data...	TO DO	ROHITROSH...
FS-11	Emotion detection		
FS-22	detect emotion from text	TO DO	S M SUT...
FS-23	store emotion that is d...	TO DO	RAHUL R...
FS-12	Text to neutral speech		
FS-24	Getting text after em...	TO DO	SOHAN M H
FS-25	Use API or GTTS...	TO DO	ROHITROSH...
FS-13	Speech modulation		
FS-26	Check the sequence of...	TO DO	S M SUT...
FS-27	Modulate the speech...	TO DO	RAHUL R...
FS-28	Generate the spee...	TO DO	ROHITROSH...
FS-14	Integrating with text editor		
FS-29	Take output of Speec...	TO DO	RAHUL R...
FS-15	Sarcasm detection [if possible]		
FS-30	Detect sarcasm with...	TO DO	SOHAN M H

Measures to be taken

- **Data quality:** The quality of the training data can have a significant impact on the performance of the model. It's important to have a large, diverse, and high-quality dataset to train the model, and to carefully preprocess and clean the data to remove any errors or inconsistencies.

- **Emotion recognition:** Emotion recognition is a challenging task in natural language processing, and it requires a deep understanding of human emotions and the ability to extract meaningful features from text. It's important to choose an appropriate approach for recognizing emotions, such as using a pre-trained emotion recognition model, or developing a custom model from scratch.

- **Speech synthesis:** Speech synthesis involves generating speech from text, and it requires a deep understanding of how speech sounds are produced and how to generate them from text. There are various approaches to speech synthesis, such as concatenative synthesis, statistical parametric synthesis, and neural synthesis, and it's important to choose the right approach for the project requirements.

- **User testing:** It's important to test the model with real users to get feedback on its performance and usability. This can be done through user studies, surveys, or other methods, and the results should be carefully analyzed and used to improve the model

References

[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, Australia saurav.pradha54@gmail.com Malka N. Halgamuge, Senior 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.

References

[3] Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep 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, doi: 10.1109/TSA.2004.828699.

References

[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

Thank You