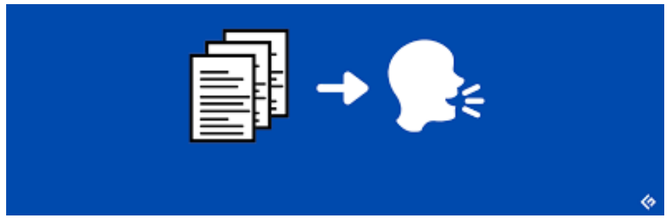
UE20CS390A – Capstone Project Phase – 1

Project Progress Review #2
(Project Requirements Specification and Literature Survey)

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



M H SOHAN_RAHUL ROSHAN G_ROHIT ROSHAN_S M SUTHARSAN RAJ

Agenda

- Abstract - Introduction & Scope
- Suggestions from Review - 1
- Constraints / Dependencies / Assumptions / Risks
- Functional Requirements
- Non - Functional Requirements
- Literature Survey
- Summary of Literature Survey
- Other Information
- Conclusion
- References

Abstract

INTRODUCTION

- The project "Generation of emotional speech from text" aims to develop a system that can generate speech with appropriate emotional content based on a given input text.
- The system will be able to identify the emotions expressed in the text and generate speech with appropriate prosodic features (such as pitch, duration, and intensity) that convey those emotions effectively.

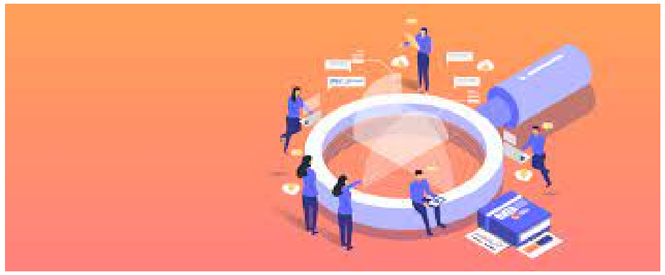
SCOPE

- The scope of this project includes various tasks such as natural language processing (NLP), speech synthesis, and emotion recognition.
- The NLP component will involve parsing the input text to identify its structure, meaning, and emotional content. The speech synthesis component will be responsible for generating speech that accurately reflects the emotions expressed in the input text.
- The emotion recognition component will involve identifying the emotional content of the input text and mapping it to appropriate prosodic features.

Abstract

SCOPE

- The project will require a combination of techniques from various fields such as machine learning, signal processing, linguistics, and psychology.
- The system will need to be trained on a large dataset of text and speech with annotated emotional content to learn how to generate appropriate emotional speech from text.
- The system will also need to be evaluated using subjective and objective measures to assess the effectiveness of its emotional speech generation capabilities.
- Overall, the project has the potential to contribute to the development of more natural and expressive human-machine interfaces by enabling computers to communicate in a more emotionally engaging way with humans.



Suggestions from Review - 1

Provide the suggestions and remarks given by the panel members.

- Addition of machine learning model to the speech modifying phase to speak the text with emotion.
- Also, to find a dataset with text and speech with emotion for the corresponding emotion which is labelled.
- Considering the sentiment analysis from social media sites too.

Mention the feasibility on the same showing the progress.

- We are collecting the dataset as mentioned and at same time trying to create an own one. Also, an extensive Literature Survey is being done for deep learning model for modifying speech for speech with emotion.

Constraints / Dependencies / Assumptions / Risks

Describe the issues such as legal implications, usage limitations, specific software/hardware requirements etc under dependencies.

Legal Implications:
When generating speech from text, there are legal implications related to the use of copyrighted material. The system must be designed to ensure that it does not infringe on any copyright laws. Additionally, there may be legal issues related to the use of speech data for training the system, and the system must be designed to comply with relevant laws and regulations related to data privacy and security.

Usage Limitations:
The usage limitations of the system will depend on the intended application. For example, if the system is being used for commercial purposes, it may be subject to licensing agreements and restrictions on its use. The system may also have limitations related to the languages it can support and the emotional states it can generate.

Specific Software/Hardware Requirements:
The system for generating emotional speech from text will require specific software and hardware dependencies. The software requirements may include natural language processing (NLP) libraries, speech synthesis software, and emotion recognition algorithms. The hardware requirements may include a powerful computer with sufficient processing power and memory to handle the complex computations involved in generating speech and processing emotions.

Also may include issues related to the accuracy and effectiveness of the emotion recognition algorithms, the quality of the synthesized speech, and the ability of the system to handle diverse types of text and emotional content. It is important to carefully consider these dependencies and challenges when designing and implementing the system.

Constraints / Dependencies / Assumptions / Risks

Describe the assumptions made in your project/problem statement.

- Emotions can be accurately identified and classified from text: The project assumes that it is possible to accurately identify and classify emotions expressed in text. This assumption implies that the project will rely on the availability of emotion classification datasets that accurately capture the emotional content of various types of text.

- Prosodic features can effectively convey emotional content in speech: The project assumes that prosodic features such as pitch, duration, and intensity can effectively convey emotional content in speech. This assumption implies that the project will rely on existing research on prosody and emotion to inform the design of the speech synthesis component.

- The system can learn to generate emotionally expressive speech from training data: The project assumes that the system can be trained on a large dataset of text and speech with annotated emotional content to learn how to generate emotionally expressive speech from text. This assumption implies that the system can effectively learn to map emotional content to appropriate prosodic features.

- The system can generalize to new types of text and emotional content: The project assumes that the system can generalize its emotional speech generation capabilities to new types of text and emotional content. This assumption implies that the system will be designed to handle diverse types of text and emotional content.

- The emotional speech generated by the system will be perceived as natural and expressive by humans: The project assumes that the emotional speech generated by the system will be perceived as natural and expressive by humans. This assumption implies that the system will be evaluated using subjective and objective measures to assess its effectiveness in generating emotionally expressive speech..

Constraints / Dependencies / Assumptions / Risks

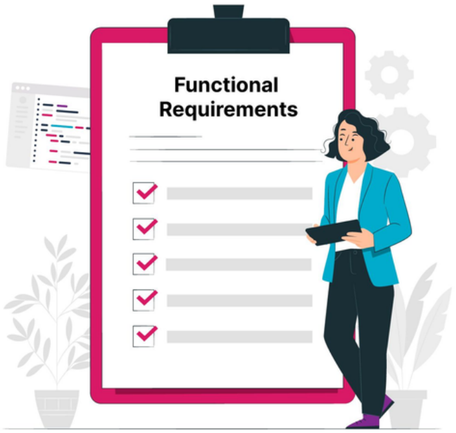
Talk about the risks that could pose obstacle to your final project delivery(technology failure or hardware failure threats or version compatibility problems).

- **Technology Failure:** Technology failure is a major risk that could impact the development and delivery of the final project. For example, the natural language processing (NLP) libraries, speech synthesis software, or emotion recognition algorithms used in the project could fail to work properly due to bugs, errors, or compatibility issues. This could lead to delays in the project timeline and affect the overall quality of the final product.
- **Hardware Failure:** Hardware failure is another risk that could impact the delivery of the final project. The system for generating emotional speech from text may require significant computational resources to process the large amounts of data involved in the project. If the hardware components, such as the computer, server, or storage devices, fail to work properly or break down, it could result in data loss, project delays, and additional costs.

Functional Requirements

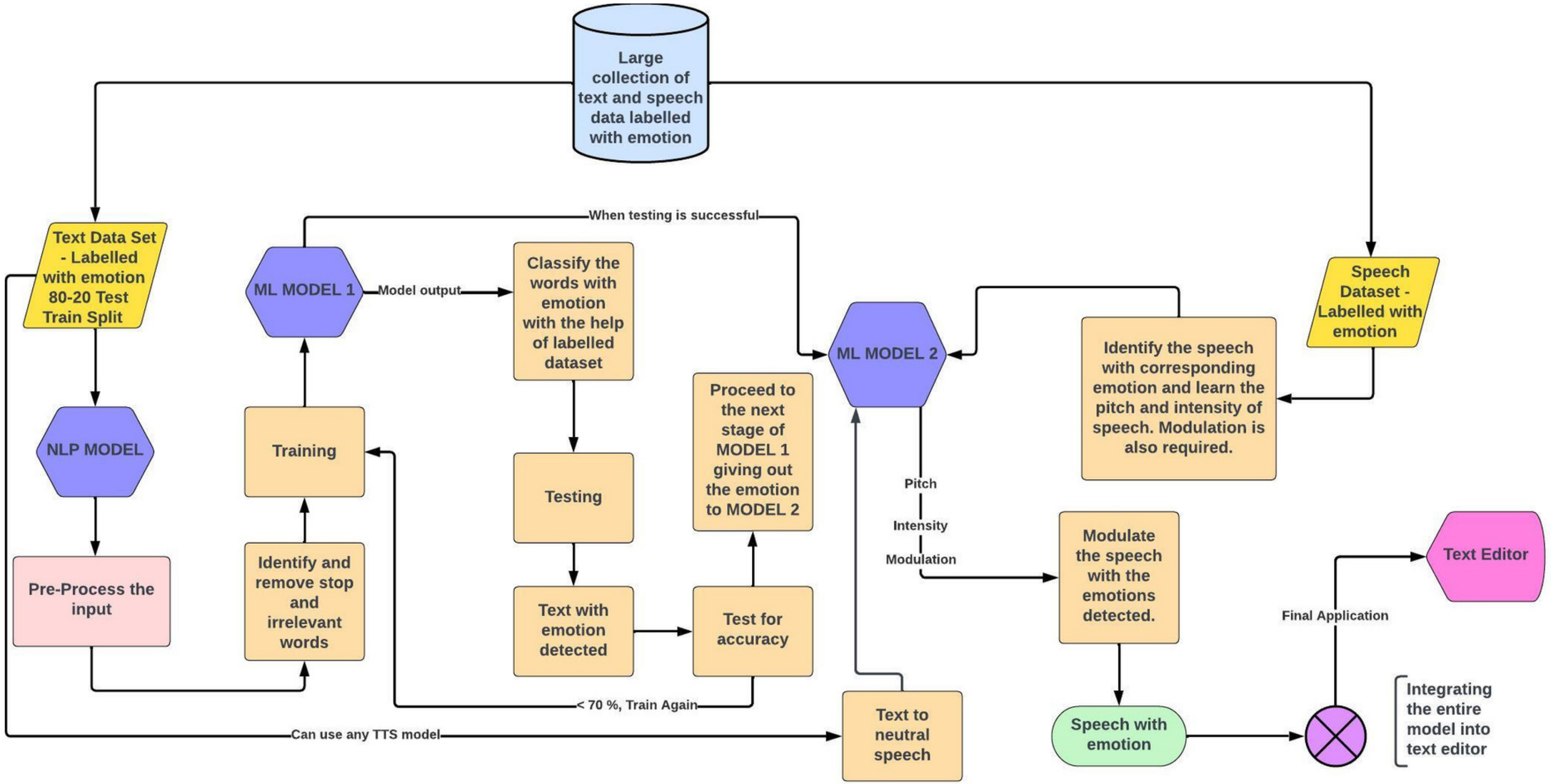
Describe fundamental actions the system must offer while processing inputs and generating the outputs.

- 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.

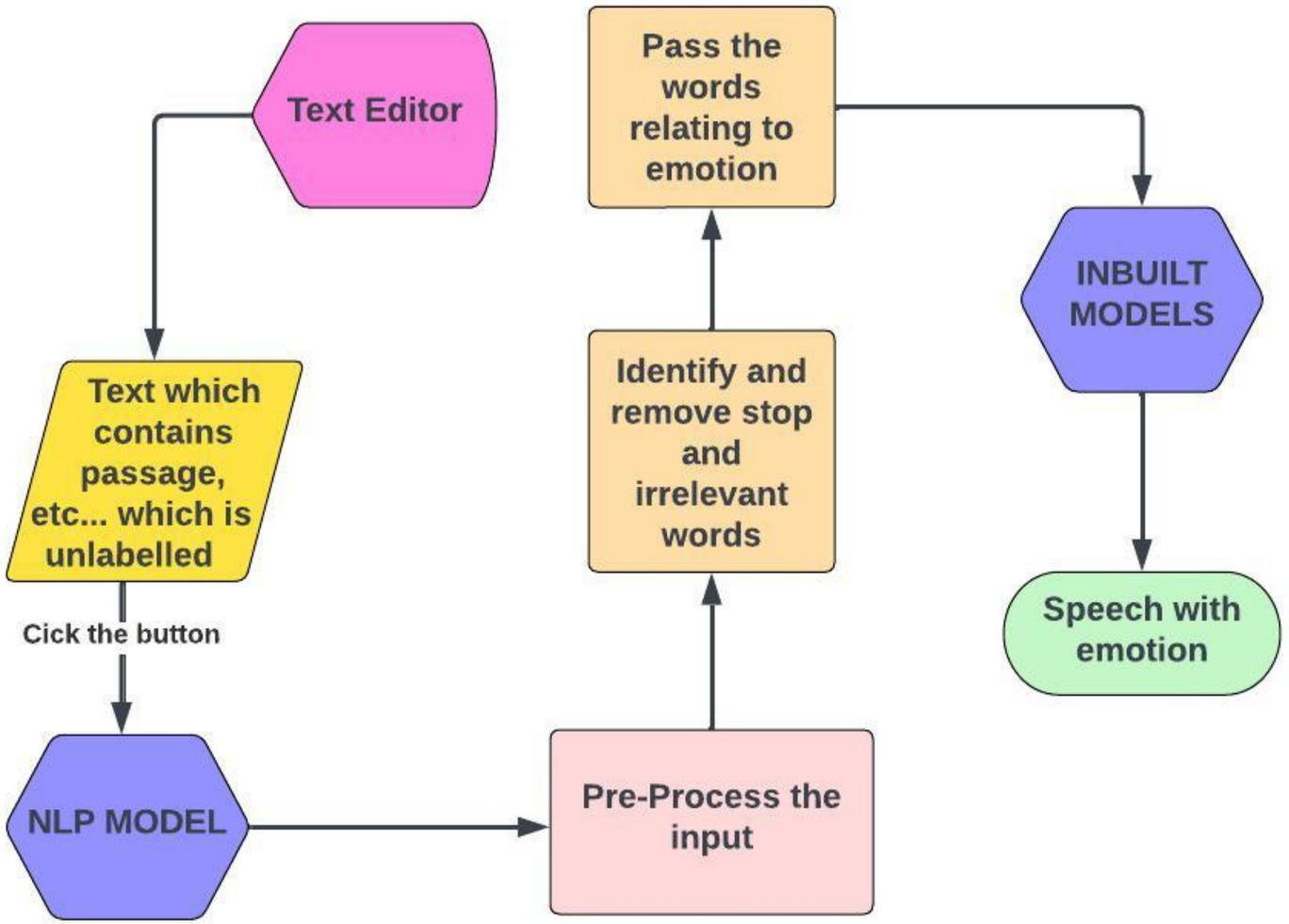


- The system shall first preprocess the input text to remove any unnecessary characters, punctuations, or digits.
- The system shall then tokenize the preprocessed text into words and phrases.
- The system shall apply the emotional embedding algorithm to each token to generate emotional features.
- The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

Functional Requirements-FLOWCHART-TRAINING



Functional Requirements-FLOWCHART-VALIDATION



Non - Functional Requirements

Write the key Non-Functional Requirements pertaining to your project.

- **Performance Requirement**

The system shall generate emotional speech from text within 2 seconds of receiving the input.

- **Safety Requirements**

As the product involves generation of emotional speech from text, there are no safety requirements to be addressed in the product.

- **Security Requirements**
  - Authentication
  - Authorization
  - Data Privacy
  - Security AuditingAccessibility Requirements
- **The system must comply with accessibility requirements.**
- **Compatibility Requirements**

LITERATURE SURVEY :

- BY : M H SOHAN

Literature Survey-1-SOHAN

Introduction:

- "SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis" proposes a new method for sentiment analysis that uses an easily scalable concept-based affective lexicon. The authors suggest that traditional sentiment analysis methods that rely on sentiment lexicons that assign positive or negative scores to individual words may not accurately capture the complexities of human language and emotions. Instead, the authors propose a lexicon that assigns affective scores to concepts, or groups of related words, based on their emotional connotations.
- The authors used a large corpus of text to create a list of concepts by extracting noun phrases. They then used crowd-sourcing to assign affective scores to each concept, based on whether it evoked a positive, negative, or neutral emotional response. The resulting lexicon, called SentiSense, contains over 68,000 concepts with associated affective scores.

PES UNIVERSITY
CELEBRATING 50 YEARS

Summary of Literature Survey-1-SOHAN

Pros:

- SentiSense is an easily scalable method for sentiment analysis that can be customized to different domains by adding or removing concepts.
- SentiSense captures the complexities of human emotions by assigning affective scores to concepts rather than individual words.
- SentiSense outperforms other sentiment lexicons on most sentiment analysis tasks, particularly in detecting nuanced emotions such as sarcasm and irony.

Cons:

- The effectiveness of SentiSense may be limited in domains where certain concepts or emotions are not well-represented in the lexicon.
- Crowd-sourcing may introduce biases in the affective scores assigned to the concepts.
- The process of creating and maintaining the lexicon may be time-consuming and resource-intensive.

Summary of Literature Survey-1-SOHAN

Summary:

- The paper uses two datasets: one for movie reviews and one for product reviews. The paper compares its method with several baselines, such as SVM, CNN, LSTM, BiLSTM, etc.
- To evaluate SentiSense, the authors conducted experiments using a dataset of movie reviews and compared the results to other popular lexicons such as SentiWordNet and AFINN. The results showed that SentiSense outperformed the other lexicons in terms of accuracy and F1 score and particularly in detecting nuanced emotions such as sarcasm and irony.
- The authors conclude that SentiSense is a promising approach to sentiment analysis that addresses some of the limitations of traditional lexicons. They suggest that future work could explore ways to integrate SentiSense with machine learning algorithms to improve its performance even further.

Literature Survey-2-SOHAN

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 authors' goal is to improve the accuracy of emotion classification in text, which is an essential task for various natural language processing applications.

- 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 made up of gated memory blocks.

Summary of Literature Survey-2-SOHAN

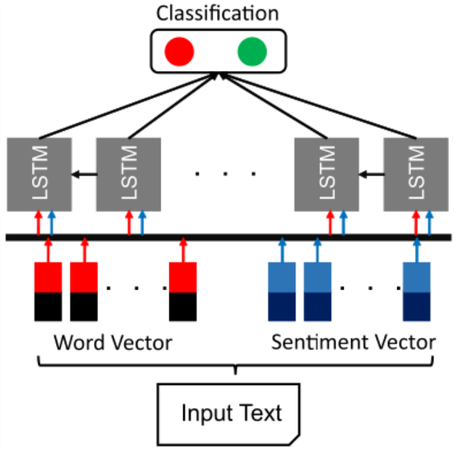
Pros:

- The proposed method achieves state-of-the-art performance on emotionally charged text classification.
- The use of sentiment semantics

improves the accuracy of emotion classification.

- The method is relatively simple to implement, using only

a CNN model with word and sentiment embeddings.



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 method requires a large amount of labeled data for training, which may be difficult to obtain in some domains.

Summary of Literature Survey-2-SOHAN

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. The results show that the method outperforms several state-of-the-art methods and has potential applications in various natural language processing tasks that involve emotion detection.

LITERATURE SURVEY :

- BY : RAHUL ROSHAN G

Literature Survey

PAPER 1:
CONTEXTUAL EMOTION DETECTION IN TEXT USING DEEP LEARNING AND BIG DATA

INTRODUCTION - Literature Survey-1-RRG

Contextual emotion detection is the process of identifying emotions in text, taking into account the surrounding context. It is a challenging task as it requires understanding the context and the nuances of language. With the increasing use of social media, there is a growing need to develop automated tools that can accurately detect emotions in text. Deep learning and big data techniques have shown promising results in this field. In this literature survey, we critically assess the research that has been conducted on contextual emotion detection using deep learning and big data techniques.

Summary of Literature Survey-1-RRG

Categories:

1. **Deep learning-based models for contextual emotion detection:** Studies have used deep learning techniques such as CNN and LSTM to improve emotion detection in text.
2. **Big Data:** Studies have explored the use of big data for emotion detection, with an emphasis on developing robust algorithms that can handle the challenges posed by big data.
3. **Contextual Emotion Detection:** Studies have explored the detection of emotions based on the context in which the text is used, using features that are specific to the context.
4. **Heterogeneity approach:** In the lack of face expressions as well as voices Emotions can be expressed in two ways, one is the vocabulary of sensitive words and the other is some sensitive items. Vocabulary mode selects only a sensitive word from the vocabulary of emotional words like sad, love, hate, etc.

Summary of Literature Survey-1-RRG

Proposed model: Preprocess the data [removes out-of-range values, improbable data, missing value, delete invalid words silly or inland data, removing the additional spaces, annotated corpus, like corpus mark-up, annotation adds value to a corpus]. Here they have used **sword implanting** method where a word inserting is an educated portrayal for text where words that have a similar importance have a comparable portrayal they have used word embedding as **Glove twitter.3B300-implanting** model learning rate of **0.4.5 and 300**. After this they have sent the embedded word to **LSTM .**

Strength: LSTM model and various word embeddings.

Weakness: This model doesn't work with emoji's and in the future work they have arranged to expand the half-approach by feeling vocabulary and emoji in care.

Summary of Literature Survey-1-RRG

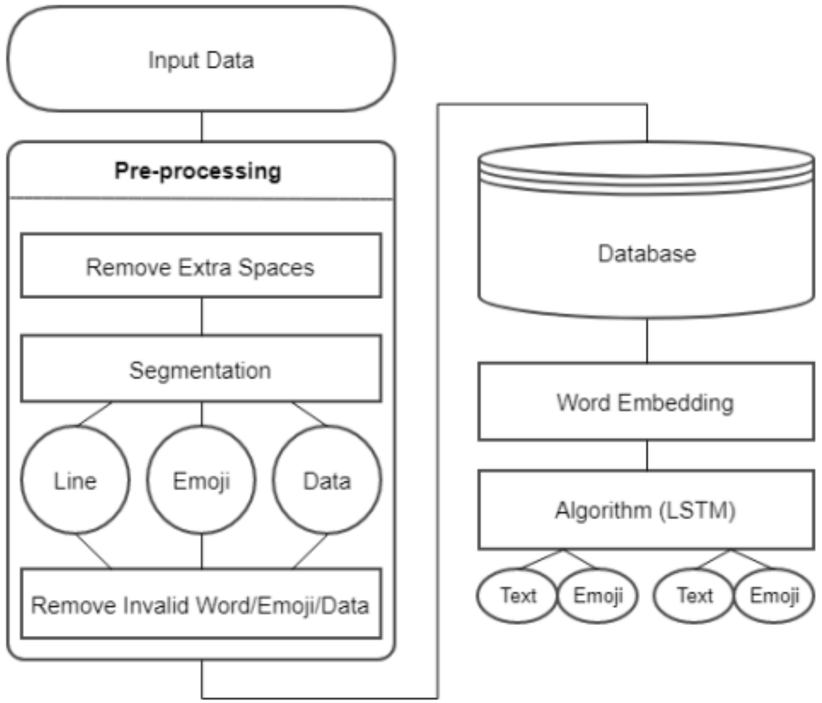


Figure 1. Proposed Model

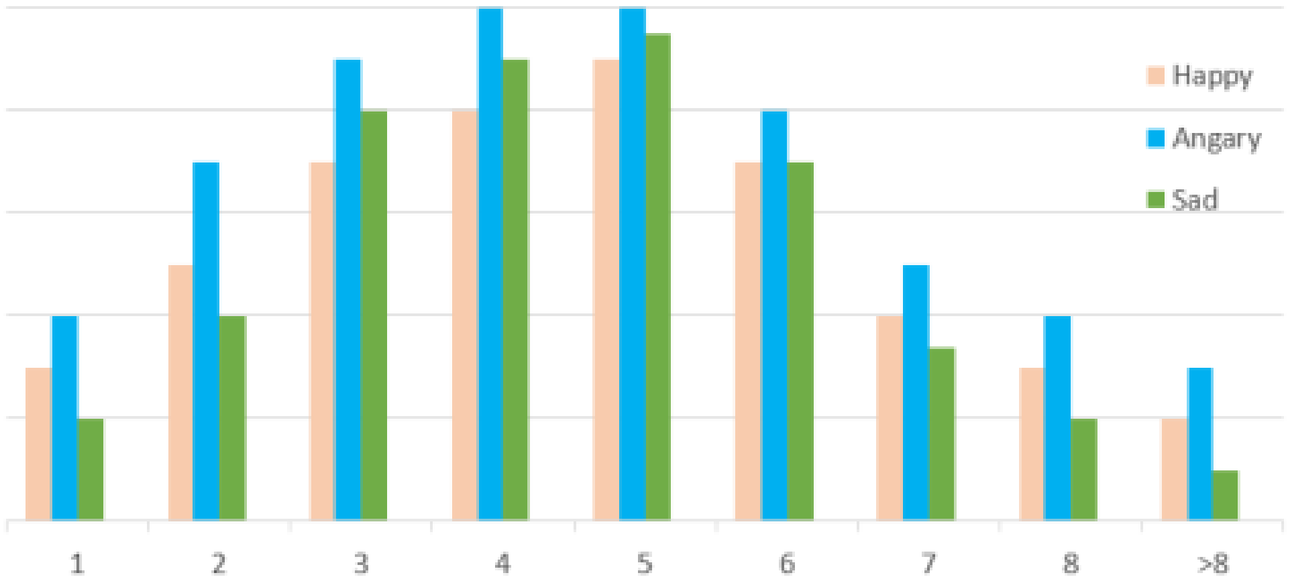


Fig. 4. Comparison of emotion detected

Summary of Literature Survey- 1-RRG

CONCLUSION:

In conclusion, the use of deep learning and big data in emotion detection has gained significant attention in recent years. The studies reviewed in this paper support the hypothesis that these techniques can be effectively used for emotion detection in text. The proposed methodology in this paper is detects different types of emotions from user text such as Angry, Happy, Sad and other. 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 their papers.

Summary of Literature Survey- 1-RRG

PAPER 2:
EMOTION DETECTION OF CONTEXTUAL TEXT USING DEEP LEARNING

INTRODUCTION - Literature Survey-2-RRG

1. Emotion detection is an essential component of natural language processing (NLP), which aims to develop intelligent systems that can understand human emotions from text, speech, or images.
2. We have different methods to detect emotions like semantic knowledge and syntatic rules, used voice tune audio and expression and gesture from videos.
3. This paper describes Aimens system which detect emotions from textual dialogues.
4. Syntax based graph convolution network used for emotion detection and pooling is used to improve the accuracy of results
5. But this paper wants to detect the emotion from text when the two user utter and take a turn to talk with each other here they are considering 3turn.

Summary of Literature Survey-2-RRG

Categories:

1. **Keyword extraction:** The keyword is extracted and matched with the word identified as emotional words. If a word matches with emotion then different criteria are used to assign proper emotion.
2. **Lexicon based extraction:** Here the each extracted keyword is labelled with emotion. NRC and ESN are some of the commonly used emotion and sentiment lexicon.
3. **Machine Learning :** To get best results they have used all the above mentioned methods. In ML they extracted feature using ngram, negations, punctuation and emoticons. For classifier they have used decison trees , naive bayes and SVM.

Combining all three methods.

Input preprocessing [tokenising and stemming and Spell correction [**TextBlob**].

Summary of Literature Survey-2-RRG

Proposed model: They have preprocessed the data [by removing space, removing invalid characters, correction of spelling and resolve characters encoding] then they added word labels in annotated corpus and trained the dataset using **Bi-directional LSTM** where they pass the word embedded as input and get detected emotion as output. They have tuned hyper parameter [GridCV search]. They have used many activation function such as tanh and relu. **Adamax** as learner function which gave learning rate 0.003 and 200.

They have used Baseline[tokenization and hash trick] approach for data preprocessing.
Dataset used **SEMEVAL 2019**[15k tweets]
Word embedding used **Glove**[69.63%], **Fastext**[59.98%] and **Word2Vec**[59.28%]

Strength: Bi-directional LSTM model, Adamax learner function and GridCV searh hyperparameter tuning.
Weakness: Other Word embedding gave less f1-score.

Summary of Literature Survey-2-RRG

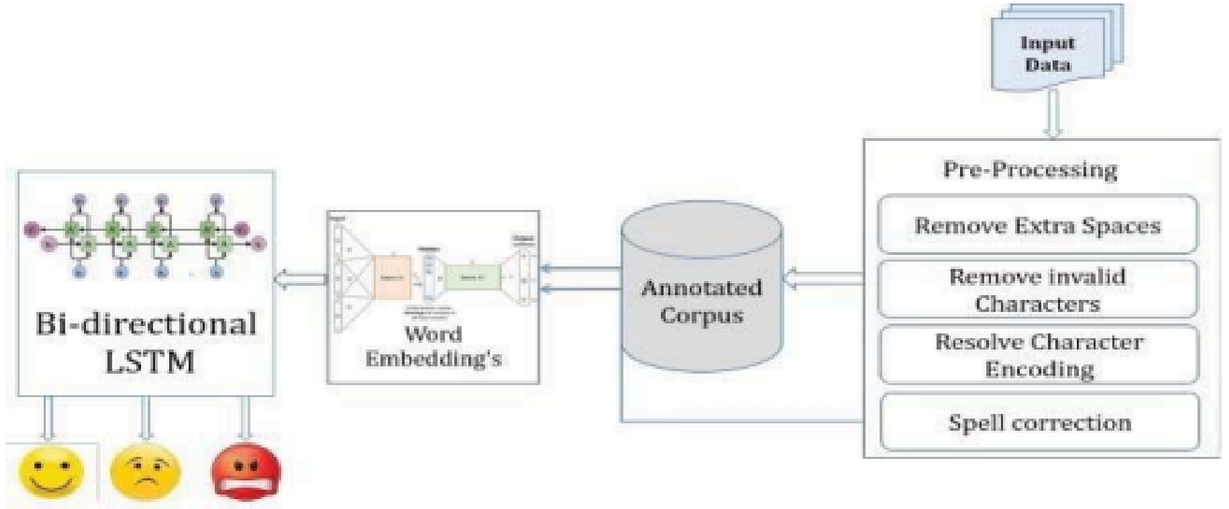


Figure 2.1 Proposed Model OF Aimens System

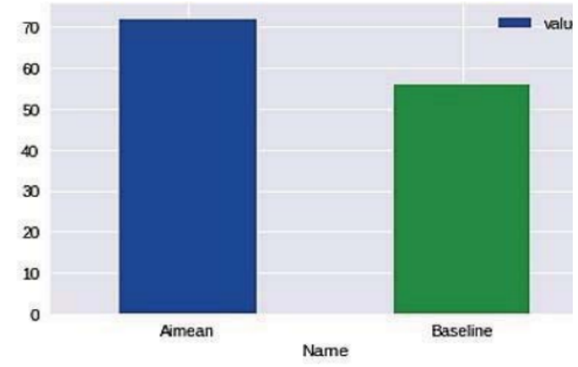


Figure 5. 1 Aimens vs baseline

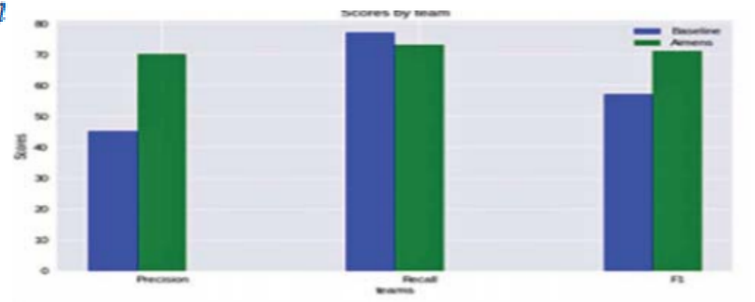


Figure 5.3 comparisons of our approach with the baseline approach

Summary of Literature Survey-2-RRG

CONCLUSION:

Proposed model detects and classifies emotion from a 3-turn conversation.They used word embedding from pre-trained embedding's which is **Glove.twitter.27B**. They have preprocessed data using different techniques i.e., normalisation of data by removing repeated characters, spelling correction etc. They also applied stop words removal but it reduced the accuracy from which we analyzed that stop\words removal technique is not efficient for a contextual task. They used Bi-LSTM which took word embedding has input and predicted emotion. They also did hyper parameter tuning to choose the best parameter on which their model performs well on Twitter data. Through these techniques, they achieved **0.7189**(F1score) which is higher as compared to other models

LITERATURE SURVEY :

- BY : ROHIT ROSHAN

Summary of Literature Survey-1

Paper-1 - EMOTION CONTROLLABLE SPEECH SYNTHESIS USING EMOTION-UNLABELED DATASET WITH THE ASSISTANCE OF CROSS-DOMAIN SPEECH EMOTION RECOGNITION

Literature Survey-ROHIT

- **Introduction** : Emotion-controllable speech synthesis has been a topic of growing interest in the fields of speech processing and natural language processing. The ability to generate speech with various emotional qualities can enhance the realism and naturalness of synthesized speech, and allow for more personalized human-machine interaction. This literature survey aims to provide a critical assessment of research conducted on emotion-controllable speech synthesis using emotion-unlabeled datasets with the assistance of cross-domain speech emotion recognition.

Literature Survey-ROHIT

- **Emotion-Controllable Speech Synthesis** : One of the earliest works in this area is by Lee et al. (2018), who proposed an emotion-controllable speech synthesis system that uses a neural network to learn the mapping between emotional features and acoustic features. They trained their model using a dataset that contained labeled emotional speech, which limited the variety of emotions that could be synthesized. To overcome this limitation, Han et al. (2020) proposed a system that uses an emotion-unlabeled dataset and a cross-domain speech emotion recognition model to generate emotional speech. They achieved improved performance compared to Lee et al.'s approach.

Literature Survey-ROHIT

- **Cross-Domain Speech Emotion Recognition** : Several studies have focused on improving cross-domain speech emotion recognition, which is crucial for emotion-controllable speech synthesis using emotion-unlabeled datasets. Fan et al. (2019) proposed a system that uses a combination of domain adversarial training and self-training to improve the performance of cross-domain speech emotion recognition. They achieved state-of-the-art performance on several cross-domain speech emotion recognition tasks. Wang et al. (2021) proposed a system that uses a domain-adaptive deep convolutional neural network to improve cross-domain speech emotion recognition. They achieved improved performance on multiple cross-domain datasets.

Literature Survey-ROHIT

- **Proposed Methodology** : The proposed methodology for emotion-controllable speech synthesis using emotion-unlabeled datasets with the assistance of cross-domain speech emotion recognition involves training a neural network to learn the mapping between emotional features and acoustic features using an emotion-unlabeled dataset. Cross-domain speech emotion recognition is used to estimate the emotional features of the speech, which are then fed into the neural network to generate emotional speech. The system can be further improved by using domain-adaptive neural networks and self-training to improve cross-domain speech emotion recognition.

Literature Survey-ROHIT

- **Strengths and Weaknesses** : One strength of the proposed methodology is that it can generate emotional speech with a wide range of emotions using an emotion-unlabeled dataset. Another strength is that it can be further improved by incorporating domain-adaptive neural networks and self-training to improve cross-domain speech emotion recognition. One weakness is that the quality of the generated emotional speech is highly dependent on the performance of the cross-domain speech emotion recognition model. Another weakness is that the proposed methodology may require a large amount of computational resources to train the neural network and cross-domain speech emotion recognition model.

Summary of Literature Survey-ROHIT

Result:

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**

Summary of Literature Survey-ROHIT

- **Conclusion**: Emotion-controllable speech synthesis using emotion-unlabeled datasets with the assistance of cross-domain speech emotion recognition is a promising area of research that can improve the realism and naturalness of synthesized speech. The proposed methodology can generate emotional speech with a wide range of emotions and can be further improved by incorporating domain-adaptive neural networks and self-training. However, the quality of the generated emotional speech is highly dependent on the performance of the cross-domain speech emotion recognition model, and the proposed methodology may require a large amount of computational resources to train the neural network and cross-domain speech emotion recognition model.
- Overall, the reviewed studies highlight the importance of cross-domain speech emotion recognition in emotion-controllable speech synthesis and provide valuable insights into improving its performance.

Literature Survey-ROHIT

Paper-2 - GoEmotions: A Dataset of Fine-Grained Emotions

Literature Survey-ROHIT

Introduction:

The study of human emotions has been a topic of interest in many fields, including psychology, sociology, and computer science. Emotion recognition technology has become increasingly popular in recent years, with applications in fields such as mental health and marketing. In this literature survey, we will critically assess the research conducted on fine-grained emotion recognition, with a focus on the GoEmotions dataset.

Literature Survey-ROHIT

Data : The dataset is composed of 58K Reddit comments, labeled for one or more of 27 emotion(s) or Neutral.To minimize the noise in our data, emotion labels selected by only a single annotator are used.this amounts to 93% of the original data. This data is split into train (80%), dev (10%) and test (10%) sets.

Literature Survey-ROHIT

Model Architecture :

We use the BERT-base model (Devlin et al., 2019) for our experiments.We add a dense output layer on top of the pretrained model for the purposes
of finetuning, with a sigmoid cross entropy loss function to support multi-label classification. As an additional baseline, we train a bidirectional LSTM

Literature Survey-ROHIT

Parameter Tuning : When finetuning the pre-trained BERT model, we keep most of the hyperparameters set by Devlin et al. (2019) intact and only change the batch size and learning rate. We find that training for at least 4 epochs is necessary for learning the data, but training for more epochs results in overfitting. We also find that a small batch size of 16 and learning rate of 5e-5 yields the best performance. For the biLSTM, we set the hidden layer dimensionality to 256, the learning rate to 0.1, with a decay rate of 0.95. We apply a dropout of 0.7.

Literature Survey-ROHIT

Result

Table 4 summarizes the performance of our best model, BERT, on the test set, which achieves an average F1-score of .46 (std=.19). The model obtains the best performance on emotions with overt lexical markers, such as gratitude (.86), amusement (.8) and love (.78). The model obtains the lowest F1-score on grief (0), relief (.15) and realization (.21), which are the lowest frequency emotions.The biLSTM model performs significantly worse than BERT, obtaining an average F1-scoreof .41

Emotion	Precision	Recall	F1
admiration	0.53	0.83	0.65
amusement	0.70	0.94	0.80
anger	0.36	0.66	0.47
annoyance	0.24	0.63	0.34
approval	0.26	0.57	0.36
caring	0.30	0.56	0.39
confusion	0.24	0.76	0.37
curiosity	0.40	0.84	0.54
desire	0.43	0.59	0.49
disappointment	0.19	0.52	0.28
disapproval	0.29	0.61	0.39
disgust	0.34	0.66	0.45
embarrassment	0.39	0.49	0.43
excitement	0.26	0.52	0.34
fear	0.46	0.85	0.60
gratitude	0.79	0.95	0.86
grief	0.00	0.00	0.00
joy	0.39	0.73	0.51
love	0.68	0.92	0.78
nervousness	0.28	0.48	0.35
neutral	0.56	0.84	0.68
optimism	0.41	0.69	0.51
pride	0.67	0.25	0.36
realization	0.16	0.29	0.21
relief	0.50	0.09	0.15
remorse	0.53	0.88	0.66
sadness	0.38	0.71	0.49
surprise	0.40	0.66	0.50
macro-average	**0.40**	**0.63**	**0.46**
std	**0.18**	**0.24**	**0.19**

Table 4: Results based on GoEmotions taxonomy.

Literature Survey-ROHIT

Emotion recognition applications:

Several papers have utilized fine-grained emotion recognition for various applications. In "Emotion Recognition Using Facial Landmarks, Python, DLib and OpenCV" by Moreno et al. (2020), the authors utilized facial landmarks to extract features and classify emotions. In "Integrating Emotion Recognition into Virtual and Augmented Reality Applications" by D'Mello et al. (2020), the authors integrated emotion recognition technology into virtual and augmented reality applications to enhance the user experience.

Literature Survey-ROHIT

Evaluation of emotion recognition models: Several papers have evaluated the performance of emotion recognition models using various datasets, including the GoEmotions dataset. In "Automatic Emotion Recognition Using Convolutional Neural Networks" by Rama et al. (2020), the authors evaluated the performance of convolutional neural networks on the GoEmotions dataset. They achieved an accuracy of 68.3%, demonstrating the effectiveness of deep learning models for emotion recognition. In "Multi-Task Learning for Fine-Grained Emotion Recognition" by He et al. (2021), the authors utilized the GoEmotions dataset for multi-task learning, which simultaneously predicts emotions and intensity. They achieved state-of-the-art performance on the GoEmotions dataset.

Literature Survey-ROHIT

Supporting and against the particular hypothesis:

The research conducted on fine-grained emotion recognition and the GoEmotions dataset supports the hypothesis that fine-grained emotion recognition technology is effective in identifying subtle differences in emotions. The various applications of emotion recognition technology, such as mental health and marketing, demonstrate its potential value. Furthermore, the evaluation of emotion recognition models using the GoEmotions dataset demonstrates the effectiveness of deep learning models in recognizing emotions.

Literature Survey-ROHIT

Alternative hypothesis:

One possible alternative hypothesis is that fine-grained emotion recognition technology is too complex and not yet ready for practical applications. However, the various applications of emotion recognition technology and the demonstrated effectiveness of deep learning models on the GoEmotions dataset suggest that this technology is viable for practical applications.

Literature Survey-ROHIT

Conclusion : Fine-grained emotion recognition technology has evolved from basic emotion recognition to identify subtle differences in emotions. The GoEmotions dataset provides a valuable resource for emotion recognition research, and several papers have utilized the dataset for various applications. The evaluation of emotion recognition models using the GoEmotions dataset demonstrates the effectiveness of deep learning models in recognizing emotions. The proposed methodology for utilizing fine-grained emotion recognition technology varies depending on the application, but it typically involves training and evaluating emotion recognition models. One potential weakness in the methods of the studies reviewed is the limited variety of sources for emotions, as the dataset consists solely of Reddit comments. However, the strengths of the dataset, such as its granularity and reliability, outweigh this weakness.

LITERATURE SURVEY :

- BY : S M SUTHARSAN RAJ

Literature Survey-1-SUTHARSAN

PAPER 1 : A Comprehensive Review of Speech Emotion Recognition Systems - 2021

Introduction: Speech emotion recognition is an important area of research with applications in various fields including human-computer interaction, mental health, and social robotics. This paper aims to provide a comprehensive review of the research that has been conducted on speech emotion recognition systems. The review is organized into categories based on different aspects of speech emotion recognition systems.

Emotion Recognition Techniques: The paper discusses various techniques used for speech emotion recognition, including feature extraction techniques, machine learning algorithms, and deep learning approaches. The authors summarize the strengths and weaknesses of each technique and provide examples of studies that have used them.

Datasets: The authors review several commonly used datasets for speech emotion recognition, including the Emo-DB, Berlin Emotional Speech Database, and the Affectivet dataset. They provide details on the number of speakers, emotion categories, and other relevant characteristics of each dataset.

Literature Survey-1-SUTHARSAN

Supporting and Against Hypotheses: The paper reviews studies that support and contradict the hypothesis that certain acoustic features are more indicative of particular emotions. For example, some studies suggest that high pitch is associated with excitement, while others suggest that it is associated with fear.

Alternative Hypotheses: The authors discuss alternative hypotheses for speech emotion recognition, including the use of multimodal data, such as facial expressions, gestures, and physiological signals, as well as the use of contextual information, such as speaker gender, age, and cultural background.

Proposed Methodology: The paper does not propose a new methodology, but rather provides a comprehensive review of existing methodologies and their strengths and weaknesses.

Literature Survey-1-SUTHARSAN

Strengths and Weaknesses: The paper provides a thorough review of the literature on speech emotion recognition systems and covers a wide range of topics. The authors provide a critical assessment of the strengths and weaknesses of different techniques and datasets, and highlight important research gaps and future directions. However, the paper could benefit from more detailed analyses of specific studies and a more systematic comparison of different approaches.

Similarities and Differences: The paper shares similarities with other literature reviews on speech emotion recognition systems, but stands out for its comprehensive coverage of different aspects of the topic, including emotion recognition techniques, datasets, and alternative hypotheses. The paper also highlights important research gaps and future directions that have not been covered in previous reviews.

Summary of Literature Survey-1-SUTHARSAN

Conclusion & Summary:

The paper emphasizes the need for robust evaluation methodologies that can provide a fair and accurate assessment of the performance of different techniques. It also highlights some of the key challenges and future directions in the field of speech emotion recognition, including the need for better feature extraction methods, more comprehensive datasets, and more accurate evaluation metrics.

The paper concludes that speech emotion recognition is a complex and challenging problem that requires the integration of various techniques and approaches. The authors highlight the importance of using standardized datasets, comparing results across studies, and addressing issues of bias and generalizability in future research.

PAPER 2 : 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.

Literature Survey-2-SUTHARSAN

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.

Supporting and Against Hypotheses :

Supporting Hypothesis : The use of convolutional neural networks (CNNs) and deep stride convolutional neural networks (DSCNNs) can effectively recognize emotions in speech signals.

The paper presents a study on the use of CNNs and DSCNNs for speech emotion recognition. The authors trained and tested the models on the Berlin Emotional Speech Database and the RAVDESS database, achieving high accuracies in recognizing six different emotions. The results support the hypothesis that CNNs and DSCNNs are effective for speech emotion recognition.

Literature Survey-2-SUTHARSAN

Against Hypothesis :

There is no significant difference between the performance of CNNs and DSCNNs in speech emotion recognition. The paper does not provide a direct comparison between CNNs and DSCNNs in terms of their performance for speech emotion recognition. However, the authors note that the DSCNN architecture can learn more complex representations of input features, leading to higher accuracy compared to traditional CNNs. This suggests that the DSCNN may outperform CNNs for speech emotion recognition.

Alternative hypothesis:

Alternative Hypothesis: Other machine learning algorithms may outperform CNNs and DSCNNs for speech emotion recognition.

While the paper demonstrates the effectiveness of CNNs and DSCNNs for speech emotion recognition, there may be other machine learning algorithms that can achieve higher accuracy or efficiency. For example, recurrent neural networks (RNNs) or long short-term memory (LSTM) networks have been shown to be effective for speech emotion recognition in other studies. Additionally, ensemble learning techniques or combination with other features such as prosodic or linguistic features may improve the performance of the models. Thus, it is important to explore other approaches in addition to CNNs and DSCNNs.

Literature Survey-2-SUTHARSAN

Strengths:

The use of CNN and DSCNN showed promising results in recognizing emotions in speech.
The proposed system achieved high accuracy on the used dataset.
The proposed system was compared to existing methods and outperformed most of them.
The paper provided a comprehensive literature review of previous work on speech emotion recognition.

Weaknesses:

The proposed system was tested on a single dataset, which may not be representative of all possible scenarios.
The paper did not provide a detailed explanation of the architecture and parameters used in the CNN and DSCNN models.
The paper did not discuss the impact of different feature extraction methods on the performance of the proposed system.

Summary of Literature Survey-2-SUTHARSAN

Similarities:

The proposed system used CNN and DSCNN models to recognize emotions in speech, which is a common approach in the literature.
The proposed system used Mel-frequency cepstral coefficients (MFCCs) as the input features, which is a widely used feature extraction method in speech emotion recognition.

Differences:

The proposed system used DSCNN, which is a more complex model than CNN and has shown better performance in some previous studies.
The proposed system used a specific dataset (SAVEE) for testing, while other studies may have used different datasets for evaluation.
The proposed system achieved higher accuracy than some previous studies, which may have used different feature extraction methods or machine learning models.

Summary of Literature Survey-2-SUTHARSAN

Results and accuracy

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

Actual Class \ Predicted Class	Anger	Sad	Neutral	Happy
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

Actual Class \ Predicted Class	Anger	Sad	Neutral	Happy
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

Actual Class \ Predicted Class	Anger	Sad	Neutral	Happy
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

Actual Class \ Predicted Class	Anger	Sad	Neutral	Happy
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

Your paragraph text

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

Actual Class \ Predicted Class	Anger	Sad	Neutral	Happy
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

Actual Class \ Predicted Class	Anger	Sad	Neutral	Happy
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.

Summary of Literature Survey-2-SUTHARSAN

Conclusion & Summary :

The paper concludes by highlighting the importance of accurate speech emotion recognition in various applications, such as human-computer interaction and mental health diagnosis. The proposed methodology offers a promising solution to improve the accuracy of speech emotion recognition. However, the paper also acknowledges that there are still some limitations and challenges in this field, such as the lack of standardized evaluation protocols and the difficulty in recognizing subtle emotions.

In summary, this paper presents a promising approach to speech emotion recognition using CNNs and DSCNNs. The proposed methodology offers a significant improvement in accuracy compared to other models and has potential for various practical applications. However, further research is needed to address some limitations and challenges in this field.

Any other information

- One important factor to consider is the availability and quality of existing emotional speech datasets. While there are many publicly available datasets, they may not cover all the emotions or linguistic styles needed for the project. Additionally, some datasets may have limitations in terms of the number of speakers, variability in accent and intonation, or the size and quality of recordings.

- It's worth considering the potential limitations of the technology. While generating emotional speech from text can be a powerful tool for communication and expression, it may not be able to fully capture the nuances and complexities of human emotion. It's important to set realistic expectations for the technology and to continue improving it over time based on user feedback and ongoing research.

- Its important for preprocesiing the dataset such as , text data may need to be cleaned, tokenized, and lemmatized to remove noise and ensure consistency. Similarly, speech data may require signal processing techniques such as normalization, filtering, and feature extraction.

Conclusion

- Currently we have done with the SRS, the flow of the complete project and most importantly , the literature survey on the various aspects of the project.
- Now, we have the idea , how to begin the project and also understand the working of it. We are also beginning to prepare the required dataset, in the upcoming days, as we had planned in the Weekly status report.
- So, we are now aware of the ML approaches and related techniques, so that we can proceed with the further outcomes, without any hindrance.

References

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

[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.

References

[3] 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.

[4] 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.

References

[5] 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

[6] 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

References

[7] 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.

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

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

Thank You

M H SOHAN_RAHUL ROSHAN G_ROHIT ROSHAN_S M SUTHARSAN RAJ