## 

## Final Year Project Report

## Emotion Recognition through voice and text using NLP techniques



## Project Advisor:

Sir Rehan Raza

## Submitted By:

Noman Nawaz (F2019266146)

Ali Salman (F2019266103)

## Session

2019-2023

## University of Management and Technology

## C-II Johar Town Lahore Pakistan

# Abstract

Emotion Recognition is a really complicated task and nowadays it is essential for us to make machines and technologies more friendly to humans for that purpose it is really being essential. We used Machine Learning (ML) concepts and models with Natural Language Processing (NLP) techniques. Emotion recognition system is an AI based technology that uses ML algorithms to detect the emotion of humans from their writings and speeches. This system collects the data of audio and converts it to text and then this data is processed and to be cleaned and this text will be predicting the emotion by using NLP techniques to that the person is happy, sad, angry, love, fear, surprise or neutral. The use of this system is diverse and will increase day by day. In this research, we have presented the overview of this system as its challenges, different applications, methodology, tools, and techniques. In this research, we have also described the advantages and limitations according to its scope. Emotion recognition through voice and text has become an increasingly popular research topic in recent years. The ability to accurately detect emotions from spoken or written language has numerous applications in fields such as psychology, healthcare, education, and marketing. This research provides an overview of the current state of the art in emotion recognition using voice and text data. It discusses various techniques and approaches used to extract emotional information from speech and text, including machine learning algorithms, natural language processing techniques, and acoustic analysis. The research also highlights the challenges faced in emotion recognition, such as the lack of labeled data and the difficulty in capturing nuanced emotions. Finally, the research discusses potential future research directions and applications for emotion recognition technology. Detecting a human emotion is a really complicated and complex task. As increasing in technology and robots we have to be strong in this field as we can and for that purpose, we have to build a model which can understand human emotion and work accordingly. There are the different applications of this system as it can be used in robotics, medicine, education, security, law enforcement, e-commerce, and social sites. ML algorithms such as SR will be used to train our model. Once our system will be able to detect the emotion of human we would be able to provide the emotional support to a person that would have to need it. NLP is a way of technique that can be used to detect or recognize the emotion of humans through text data. This system will first have the text or speech data in it and if it is audio it will be converted into text by using the SR algorithm. This will be our main entity for this to convert the audio or speeches into text and first, we have to remove the noise and resampling of audios so that it can be cleared and then it will go for the classifiers. We selected the datasets of different emotions for our module and trained them for the maximum accuracy. We used the text and audio datasets to train this module. Writings and speeches are the main fundamentals of people communicating or expressing their emotions and thoughts. And detecting the emotions of one would be the best approach for understanding human nature and their responses to surroundings. It will predict the emotion of a human as they are happy, sad, angry, love, fear, surprise, or neutral.

Table of Contents

[Final Year Project Report i](#_Toc138250169)

[Emotion Recognition through voice and text using NLP techniques i](#_Toc138250170)

[Project Advisor: i](#_Toc138250171)

[Submitted By: i](#_Toc138250172)

[Session i](#_Toc138250173)

[University of Management and Technology i](#_Toc138250174)

[C-II Johar Town Lahore Pakistan i](#_Toc138250175)

[Dedication ii](#_Toc138250176)

[Final Approval iii](#_Toc138250177)

[Acknowledgements, iv](#_Toc138250178)

[Abstract v](#_Toc138250179)

[Definitions and Acronyms viii](#_Toc138250180)

[1. Introduction 1](#_Toc138250181)

[1.1 Problem Overview 1](#_Toc138250182)

[1.2 Research Questions 2](#_Toc138250183)

[1.3 Research Objectives 2](#_Toc138250184)

[1.4 Scope 2](#_Toc138250185)

[1.5 Methodology 2](#_Toc138250186)

[1.6 Significance/ Potential Applications 3](#_Toc138250187)

[2. Background 5](#_Toc138250188)

[3. Literature review 7](#_Toc138250189)

[3.1 Detailed Literature Review 11](#_Toc138250190)

[3.2 Gap Analysis 16](#_Toc138250191)

[4. Proposed Methodology 17](#_Toc138250192)

[4.1 Suggested Approach 17](#_Toc138250193)

[4.2 Workflow of the system 18](#_Toc138250194)

[4.3 Algorithms/Architecture 19](#_Toc138250195)

[5. Design and Implementation 20](#_Toc138250196)

[5.1 System Design 20](#_Toc138250197)

[5.2 System Implementation 20](#_Toc138250198)

[5.3 Assumptions/Constraints (Optional) 21](#_Toc138250199)

[6. Evaluation 22](#_Toc138250200)

[6.1 Experimentation 22](#_Toc138250201)

[6.2 Experimental Setup 22](#_Toc138250202)

[6.3 Experiments Design/Details 22](#_Toc138250203)

[6.4 Discussion/Analysis 23](#_Toc138250204)

[6.4.1 Dataset: 23](#_Toc138250205)

[6.4.2 Models: 26](#_Toc138250206)

[6.4.2.1 Logistic Regression: 26](#_Toc138250207)

[6.4.2.2 Naïve Bayes: 28](#_Toc138250208)

[6.4.2.3 Decision Tree: 29](#_Toc138250209)

[6.4.2.4 Random Forest: 31](#_Toc138250210)

[6.4.3 Results: 32](#_Toc138250211)

[6.4.4 Evaluation: 32](#_Toc138250212)

[6.4.5 Comparative Analysis: 33](#_Toc138250213)

[6.4.5.1 Different Researches and their proposed models with methodology and results: 33](#_Toc138250214)

[6.4.5 Limitations and challenges: 34](#_Toc138250215)

[6.4.6 Insights: 34](#_Toc138250216)

[7. Conclusion and Future work 35](#_Toc138250217)

[8. References/ Bibliography 36](#_Toc138250218)

[9. Appendix 38](#_Toc138250219)

[9.1 Glossary of terms 38](#_Toc138250220)

[9.2 Pre-requisites 38](#_Toc138250221)

[9.3 Reference/ Source Documents 39](#_Toc138250222)

Table of figures

[Figure 1 Suggested Approach 25](#_Toc138249167)

[Figure 2 Workflow of the system 26](#_Toc138249168)

[Figure 3 Logistic Regression 35](#_Toc138249169)

[Figure 4 Naive Bayes 36](#_Toc138249170)

[Figure 5 Decision Tree 38](#_Toc138249171)

[Figure 6 Random Forest 39](#_Toc138249172)

# 1. Introduction

In the era of technology where we are making automotive vehicles as part of our lives and with the evolution of automotive vehicles getting towards autonomy, the integration of Machine Learning (ML) and Artificial Intelligence (Ai) powered systems stands as both a breakthrough but also a vulnerability. The automotive innovation has been relies on the technologies, promising safer, more comfort, efficient and intelligent transportation solutions [1]. As this journey is increasing day by day towards the modification of technologies with vehicles it also concerns about the big impact of interplay between attack and defense within these systems and specifically in the contect for safety and reliability of humans [2].

This automotive industry’s stride directly towards promising convenience and improving safety, by implementing the Machine Learning (ML) algortihms. [3]. The most innovative design forward present conundrum- a paradoxical vulnerbality to malicious attacks. As these vcehicles becomes more interconnected, relying on the sophisticated algorithms for decision making and functionality, they are becoming susceptible of cyber threats and adversarial manipulations[4].

Incidents happening recently and further study in this domain have brough to the light of the vulnerbalities we will be having with the AI- powered systems for the automotive [2]. The high- profile attacks on the vehiclesm where adversarial inputs have led to compromised functionality and with the safety. It servces as potent reminders of threat is landscape [1]. The ramifications of suc attacks are not just theroritical but have tangible concequences too and potentionally to humans and to also for the trust worthiness of vehicles [4].

This report ains to explore the attack and defence in ML and Ai powered automtovie systems. The motivation behind is to understand and address the vulnerabilities while harnessing the potential of these techonlohgies colliding with automotive vehicles. By devling into this balance, we have got many of the complexities surrounding with the references of security and relialblity of automonous vehicles.

So there would be the robust defence mechanisms against the adversarial threats we are having from collioson of automotive vehcilces with Ai and ML. This will be the exact solution for reaching out and maximisng useage of this techonology but for the betterment and also with the safety precautions.

The goverenment’s interest in regulation and ensuring the safrty of automonus vehciles adds is the importance of this discussion that how major impact this has on humans. There must be some policy frameworks and regulatory to make sure following the guidelines and must be informed by understanding of what vulnerabilities these techonologies are bringing in the humans lives and how badly this gonna effect

In response to the challenges been facing this research of the report will focus on the depth analysis fo attack vectors, defencse strategis, and proposes innovation solutions aiming at forigying the Ai powered automotive systems. Drawing insights from academic research, industry reports and governmental initiatives, this report will aims to contiribute the evolving od securing the future automous vehicles and humans.

## Problem Overview

The integration of ML and Ai powered systems in autonomous vehicles introduces many of the vulnerabilities that expose them to major adversarial attakcs. The attacks can compromise the functionality, safety, reliability and trust in autonomous vehicles for humans. Highlighting this problem would need for roubust defecne mechanisms to safeguward against such threats.

## Research Questions

* What are the primary attacks vectors that target ML models in autonomous vehciles?
* What defecne strategies must be used for effectively to reduce the adversarial threats in Ai powered automotive system?
* What are the things to be considered before colliding these vehicles with Ai?

## Research Objectives

* Identify and analyze the attack vectors on ML models used in autonomous vehciles.
* Investigate exisiting defence mechanisms and their effectiveness.
* Propose the innovative solutions aimed for better security of vehicles against cyber attacks

## Scope

This research will specifically focus on the interplay between the attack and defense in ML and Ai powered automotive systems. It aims to analyze the vulnerabilities, mainly to adversarial attacks as cyber attacks, and proposing the strategies and exact solutions to these threats. The study of this will also focus on policy frameworks and guidelines needs to be considered for this.

## Methodology

The research methodology encompasses a comprehensive review of academic literature, industry reports, and governmental policies related to AI-powered automotive systems' security vulnerabilities. The analysis involves identifying attack vectors through case studies, evaluating existing defense mechanisms, and proposing innovative solutions through empirical simulations and theoretical models [2, 4]. The study will employ a mixed-method approach, combining qualitative and quantitative analyses to achieve comprehensive insights into the intricacies of attack-defense dynamics in autonomous vehicles.

1. **Adversarial Loss Function (for adversarial training):**

* Adversarial training of AI models in the defense strategies against attacks

Where:

* + *θ represents model parameters.*
  + *x denotes the input.*
  + *y is the true label.*
  + *δ stands for the perturbation.*
  + *S is the allowable perturbation set.*
  + *ℓℓ is the base loss function.*

1. **Robust Optimization Objective:**

* This equation represents the objective function used in training models for adversarial robustness.

1. **Robustness Metric (for model evaluation):**

*N1​∑i=1N​I(f(xi​+δi​;θ)=yi​)*

Where:

* N is the total number of samples.
* xi​ is the input sample.
* δi​ is the perturbation applied to the input.
* yi​ is the true label.
* θ represents the model parameters.
* f is the model's prediction function.
* I is the indicator function.

# analysis and discussion

Understanding Adversarial Threat Vectors

1. Adversarial Threat Landscape:

• Threat Taxonomy: Categorizing adversarial threats based on attack vectors, exploring how each type of attack targets specific components of autonomous vehicle ML systems.

• Attack Methodologies: Detailing the technical intricacies of adversarial attacks, examining the mechanisms behind perturbations in sensor data, poisoning of training datasets, and manipulations in environmental cues like road signs or markings.

2. Complexity of Attacks:

• Impact Assessment: Evaluating the potential impact of successful adversarial attacks on the decision-making processes of autonomous vehicles, considering scenarios where compromised ML models misinterpret inputs in real-time driving conditions.

• Unforeseen Threats: Discussing the challenges posed by novel and evolving adversarial tactics, highlighting the necessity for adaptive defense mechanisms to counter unforeseen threats

Understanding Adversarial Threat Vectors

1. Adversarial Threats

* Threat Texonomy:

Categorizing the threats specifically the adversarial threats and their attack vectors, exploring how each type of attack targets specific components of autonomous vehicle ML systems

* Attack Metholodogy

Examining the technical intricacies of the threats, attacks and the mechanisim behind every attack or threat in the sensor data, directing from datasets trained models, and manipulations in the environment from our surrounding things.

2. Complexity of Attacks:

* Impact Assesments:

Defining that what are the major impacts of adversarial attacks happened on the decision making processes of autonomous vehicles,

# Literature review

There are many kinds of research and approaches that have been made separately for predicting the emotion of humans through writing and audio separately. But in combination, there is nothing like a system that can predict emotion through voice and text collectively or at the same time.

We have read many of the research papers in numbers of 40 to 50 with different technologies. We compare all the concepts and algorithms and finally create a better hybrid system having both functionalities.

Table 3 Research Papers

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Ref. #*** | ***Problem*** | ***Methodology*** | ***Datasets*** | ***No. of Classes*** | ***Pre-Processing*** | ***Model*** | ***Evaluation*** | ***Results*** | ***Strength and Weaknesses*** |
| ***1*** | *Emotion detection from text and speech* | *GMMs, SVM, HMM, Neural Network* | *EmoBank, ISEAR, EmoLex* | *2* | *ML, Naïve Bayes, BWS Techniques* | *CNN joint learning* | *Ekman's Model* | *85% Accuracy* | *Good datsets be used but not the latest technologies* |
| ***2*** | *Speech Emotion Recognition Using Deep Learning Techniques* | *DL and SER* | *eNTERFAC' 05 and EMOD database* | *7* | *HMM, DNN, ANN, SER techniques* | *RNN* | *Database utilised and performance* | *61 % Accuracy* | *Efficient models but least accuracy* |
| ***3*** | *state-of-the-art approaches for emotion recognition in text* | *DL, Hybrid* | *ISEAR, SemEval 2007, 2008, 2017* | *4* | *NLP, Keyword-spotting techniques* | *SVM* | *Multi label accuracy* | *67 % Accuracy* | *The use of semantic features but classes gets confused* |
| ***4*** | *Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM* | *SER* | *IEMOCAP, Emo-DB, RAVDESS* | *4* | *K-mean, RBF, STFT Algorithm* | *CNN, SER* | *BiLSTM, K-mean, FC-1000* | *72 % Accuracy* | *The architecture can be used for future purposes but not good models* |
| ***5*** | *Speech Emotion Recognition with Multi-task Learning* | *GMMs, SVM, HMM, MLP* | *Emiratu-accented and SUSAS* | *6* | *SVM and MLP* | *DNN-GMM* | *GMM-DNN* | *83.97 % Accuracy* | *Good accuracy and models but not good architectures* |
| ***6*** | *Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining* | *Hybrid, DL* | *USC-IEMOCAP* | *4* | *SVM* | *UBM,GMM* | *eVECTOR and BoW features* | *69.2 % Accuracy* | *Suitable models but not better performance in acoustic* |
| ***7*** | *Deep learning approach to text analysis for human emotion detection from big data* | *Hybrid, SVM, NLP* | *SemEval-2007, Enteraface'05 Emotion* | *5* | *Sentence Splitter, NLTK* | *SVM* | *POS tagging, NLP* | *90 % Accuracy* | *Module can be implement as input to devices but not good architectures* |
| ***8*** | *SPEECH EMOTION RECOGNITION USING SELF-SUPERVISED FEATURES* | *NLP,DL* | *IEMOCAP* | *8* | *NLTK* | *DLSTA* | *NLP techniques* | *97.22 % Accuracy* | *Use of latest technology and architecture but not good datasets* |
| ***9*** | *Multiomdel Speech emotion recognition using audio and text* | *DNN, RNN* | *IEMOCAP* | *4* | *NLP, RNN* | *ASR, TRE* | *WAP, NLP techniques* | *43.51 % Accuracy* | *Latest models developed but not suitable datasets is used* |
| ***10*** | *Detecting Emotion from Natural Language text using hybrid and NLP models* | *NLP, ML,DL* | *Online social Network Communication (Kaggle)* | *5* | *Vectorinzation, tokenize* | *Bi-LSTM, Hybrid* | *Naive Bayes, Bi-LSTM and Hybrid* | *61 % Accuracy* | *Use of latest technology and datasets but not better performance* |
| ***11*** | *Multiomodel emotion recognition with high level speech and text features* | *CNN, Word2vec and GloVe* | *IEMOCAP, LibriVox* | *4* | *BLSTM, LSTM* | *CNN, SER,TER* | *Wav2vec 2.0* | *70.1 % Accuracy* | *Use of high level features and classifiers but only using pre trained models* |
| ***12*** | *Bimodal speech emotion recognition using pre trained language models* | *LSTM, SER, ASR* | *IEMOCAP* | *4* | *ASR, SER, TER combining* | *SER based XLNET-Base* | *BERT and XLNet* | *69.5 % Accuracy* | *Combinig all the models and using best features but still less accuracy* |
| ***13*** | *Deep neural networks for emotion recognition combining audo and transcripts* | *LSTM* | *IEMOCAP* | *4* | *ASR, DNN* | *LSTM, GRU networks* | *MCNN* | *68.8 % Accuracy* | *Using pre trained models in a good way but not good architectures is used* |
| ***14*** | *Efficient speech emotion recognition using multi scale cnn and attention* | *Bi-RNN, MSCNN, ASR* | *IEMOCAP* | *4* | *NLP, SPU* | *MSCNN-SPU* | *MFCC, X-Vector and TDNN* | *74 % Accuracy* | *Good architecure diagrams but nothing new in model only few changes* |
| ***15*** | *Multimodel Speech emotion recognition using audio and text* | *NLP, RNN, NLTK* | *IEMOCAP* | *4* | *DNN, CNN, ARE, TRE* | *Multimodel MDRE* | *ARE, MFCC, TRE, WAP* | *70 % Accuracy* | *New Multimodel but accuracies is reduced in before introduced models* |
| ***16*** | *Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining* | *Matlab frequency, SVM classifier.NLP* | *SemEval-2007* | *6* | *NLTK* | *Hybrid SVM* | *MFCCs* | *81 % Accuracy* | *Hybrid model but older techniques and pre trained* |
| ***17*** | *An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network* | *CNN, Ekman's Emotion model,NLP* | *SemEval-2013* | *7* | *NLTK* | *CNN* | *RF, LSTM* | *80.99 % Accuracy* | *Good approach and accuracy but not good architetcures* |
| ***18*** | *MULTI-HEAD ATTENTION FOR SPEECH EMOTION RECOGNITION WITH AUXILIARY LEARNING OF GENDER RECOGNITION* | *SER, NLP* | *IEMOCAP* | *4* | *NLTK* | *SER based emotion* | *SER* | *76 % Accuracy* | *Gender recognition but Limited to SER only not good architetcures* |
| ***19*** | *Speech Emotion Recognition Using Speech Feature and Word Embedding* | *NLP, LSTM* | *IEMOCAP* | *4* | *SER, MFCC, NLTK* | *LSTM, CNN* | *SER, SVM* | *75.49 % Accuracy* | *Good pre trained model used but not good architetcures* |
| ***20*** | *FUSION APPROACHES FOR EMOTION RECOGNITION FROM SPEECH USING ACOUSTIC AND TEXT-BASED FEATURES* | *SER, SVM* | *MSP-PODCAST* | *4* | *BERT, MFCC* | *Fusion model* | *BERT* | *70 % Accuracy* | *New model and good dataset used but not good architectures* |
| ***21*** | *Audio-Textual Emotion Recognition Based on Improved Neural Networks* | *Bi-LSTM, DNN* | *IEMOCAP* | *4* | *MFCC, LSTM* | *CNN, LSTM* | *UA, WA on Matrices* | *Improved 13 %* | *Good accracy and architectures but bad Evaluation* |
| ***22*** | *DEEP MULTIMODAL LEARNING FOR EMOTION RECOGNITION IN SPOKEN LANGUAGE* | *CNN, LSTM* | *IEMOCAP* | *5* | *NLTK,POS, MFCC* | *CNN-LSTM, and DNN* | *Multimodel CNN-LSTM based* | *60.4 % Accuracy* | *Good model and architecture but badly trained* |
| ***23*** | *SPEECH EMOTION RECOGNITION USING MULTI-HOP ATTENTION MECHANISM* | *NLP, BRE* | *IEMOCAP* | *4* | *NLTK, BLSTM* | *DNN based MHA 1-2* | *E vec-MCNN-LSTM* | *6.5 % Improved* | *Good techniques and model but architectures* |
| ***24*** | *Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network* | *RNN, SER* | *IEMOCAP* | *4* | *MFCC, BERT* | *CMU-MOSEI* | *SER* | *9.81 % Improved* | *Good techniques, model and architectures but weak in evaluation* |
| ***25*** | *A Bimodal Approach for Speech Emotion Recognition using Audio and Text* | *SVM, Word2Vec, FastText and BERT* | *RAMAS* | *7* | *NLP* | *Bimodal fusion* | *LR, RF, NB* | *60% Accuracy* | *Good models and techniques but bad architectures and classifier* |
| ***26*** | *MULTI-MODAL EMOTION RECOGNITION ON IEMOCAP WITH NEURAL NETWORKS* | *LSTM* | *IEMOCAP* | *10* | *MFCC* | *Mocap Model1* | *MLP, LSTM* | *50 % Accuracy* | *Pre trained good models combined but least accuracy* |
| ***27*** | *Towards improving e-commerce customer review analysis for sentiment detection* | *NLP* | *SQUAD* | *4* | *NLTK, BLSTM* | *CNN based NMT* | *BERT, GLOVE, ELMo* | *79.83 % Accuracy* | *Good processing, architectures but limited scope* |
| ***28*** | *Multimodal Embeddings From Language Models for Emotion Recognition in the Wild* | *NLP, SER, DNN* | *CMU-MOSEI* | *4* | *LSTM* | *BiLM* | *UAR, SER* | *60 % Accuracy* | *Good models and techniques but bad architectures and classifier* |
| ***39*** | *Speech Emotion Recognition Using Spectrogram & Phoneme Embedding* | *CNN,DNN, ASR* | *USC-IEMOCAP* | *4* | *HMM,SVM, GMM* | *CNN based STFT* | *CNN Model 3* | *62.8 % Accuracy* | *Good architectures and pre trained model but not classifier* |
| ***30*** | *Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations* | *NLP* | *CMU-MOSEI* | *6* | *MFCC* | *UniSpeech-SAT* | *NLTK* | *68 % Accuracy* | *Good architectures and classifires but complicated model* |

## 3.1 Detailed Literature Review

Here is the detailed literature review of 30 papers mentioned in the table:

1. **Emotion detection from text and speech**

This paper provides an extensive review of the current research and techniques employed in the detection of emotions from both text and speech data. The authors provide a detailed description of emotions and their importance in communication, followed by a comprehensive overview of the techniques used in the field, including supervised and unsupervised machine learning algorithms, sentiment analysis, and feature extraction techniques. The paper discusses the challenges and limitations of existing approaches and identifies opportunities for future research.

.

1. **Speech Emotion Recognition Using Deep Learning Techniques**

The article starts by providing a brief overview of the importance of speech emotion recognition and its applications in various fields such as psychology, medicine, and human-computer interaction. They categorize the reviewed papers based on the type of deep learning techniques used, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Hybrid Neural Networks.

1. **State-of-the-art approaches for emotion recognition in a text**

The authors start by introducing the concept of emotion recognition in text and its importance in various fields such as psychology, marketing, and social media analysis. They also mention the challenges associated with this task, such as the ambiguity of emotional expressions, the cultural and linguistic variations, and the lack of labeled datasets.

The paper then proceeds to review the existing approaches for emotion recognition in text, starting with the traditional rule-based methods that use predefined lexicons and linguistic rules to identify emotions.

1. **Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM**

This paper proposes a clustering-based approach for speech emotion recognition. The authors incorporate learned features and a deep BiLSTM (Bidirectional Long Short-Term Memory) network to improve the performance of the proposed approach.

The paper begins with an introduction to the importance of speech emotion recognition and the challenges associated with it. The authors then review existing techniques for speech emotion recognition, including feature extraction, classification, and deep learning-based methods. They highlight the limitations of these methods and the need for a more effective approach.

The proposed clustering-based approach involves first clustering the input speech signal into multiple clusters, each representing a specific emotion.

1. **Speech Emotion Recognition with Multi-task Learning**

This paper by Ismail Shahin, Ali Bou Nassif, and Shibani Hamsa presents a novel approach for emotion recognition using a combination of a Hybrid Gaussian Mixture Model (HGMM) and Deep Neural Network (DNN) architectures. The study is motivated by the need for robust and efficient emotion recognition systems, which can have applications in various fields, including human-computer interaction, healthcare, and entertainment. They introduce their proposed method, which combines the strengths of HGMM and DNN. The output of the HGMM is then fed into a DNN, which is trained using backpropagation to classify the input into one of the target emotion categories. The authors use a dataset of Arabic speech signals to evaluate their proposed approach, achieving an accuracy of 82.5%, which outperforms several state-of-the-art approaches.

1. **Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining**

The literature review of the paper provides an overview of previous studies on emotion classification, particularly in the context of audio conversations. The paper discusses various approaches that have been proposed in the literature, including rule-based systems, statistical models, and machine-learning algorithms. The authors highlight the limitations of these approaches, particularly in the context of emotion classification of audio conversations. The paper emphasizes the importance of combining multiple modalities, such as text and speech, for accurate and robust emotion classification.

1. **Deep learning approach to text analysis for human emotion detection from big data**

The study aims to address the challenge of detecting human emotions from big data using text analysis techniques. To achieve this, the author proposes a deep learning approach that involves the use of a neural network model known as Long Short-Term Memory (LSTM). The LSTM model is trained using a large dataset of human emotions expressed in textual data.

The literature review of the paper reveals that there is a growing interest in the use of deep learning techniques for natural language processing tasks, including sentiment analysis and emotion detection.

.

1. **SPEECH EMOTION RECOGNITION USING SELF-SUPERVISED FEATURES**

The authors highlight the limitations of these approaches, particularly in terms of the need for large amounts of labeled data and the generalization performance of the models. The paper emphasizes the importance of developing methods that can effectively learn from unlabelled data for SER.

1. **Multiomdel Speech emotion recognition using audio and text**

The authors note that traditional approaches to SER have limitations, particularly in terms of handling noisy data and requiring large amounts of labeled data. They also note that the use of deep learning and multimodal fusion has shown promise in improving the performance of SER models.

The authors then present their proposed approach for multimodal SER, which involves using a deep neural network architecture that combines audio and text features.

1. **Detecting Emotion from Natural Language text using hybrid and NLP Models**

The authors then discuss the different approaches that have been used to detect emotions in text, including rule-based approaches, machine learning-based approaches, and deep learning-based approaches. They provide an overview of the advantages and limitations of each approach, highlighting that deep learning-based approaches have shown promising results in recent years.

The paper then introduces the concept of pre-trained NLP models, which are trained on large datasets and can be fine-tuned for specific tasks such as emotion detection.

1. **Multimodel emotion recognition with high-level speech and text features**

The authors describe the advantages and limitations of different modalities, emphasizing that speech and text are particularly important for capturing complex emotions.

The authors then discuss the different approaches that have been used for multimodal emotion recognition, including feature-level fusion, decision-level fusion, and hybrid approaches. They provide an overview of the advantages and limitations of each approach, highlighting that hybrid approaches have shown promising results in recent years.

1. **Bimodal speech emotion recognition using pre-trained language models**

The authors present a comprehensive literature review of existing research on speech emotion recognition and highlight the challenges and opportunities in this field.

The paper then introduces the concept of pre-trained language models, which are trained on large amounts of text data and can be fine-tuned for specific tasks such as speech emotion recognition. The paper then presents a bimodal speech emotion recognition system that combines pre-trained language models with acoustic features.

1. **Deep neural networks for emotion recognition combining audio and transcripts**

The study aims to improve the performance of emotion recognition systems by combining audio and text data using deep neural networks. The authors begin with a comprehensive review of existing literature on emotion recognition, highlighting the challenges and opportunities of multimodal emotion recognition. They note that emotions are conveyed through different modalities, such as speech, text, and facial expressions, and combining these modalities can enhance the accuracy and robustness of emotion recognition systems.

The paper then presents a multimodal emotion recognition system that combines audio and text data using deep neural networks. Specifically, the authors use convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to extract relevant features from audio and text data, respectively. The extracted features are then combined using a hybrid CNN-LSTM network.

1. **Multimodel Speech emotion recognition using audio and text**

The study aims to improve the performance of speech emotion recognition systems by combining audio and text data using a multimodal approach. The authors begin with a comprehensive review of existing literature on emotion recognition, highlighting the challenges and opportunities of multimodal emotion recognition. They note that emotions are conveyed through different modalities, such as speech, text, and facial expressions, and combining these modalities can enhance the accuracy and robustness of emotion recognition systems.

The paper then presents a multimodal speech emotion recognition system that combines audio and text data using a combination of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks..

1. **Efficient speech emotion recognition using multi-scale cnn and attention**

The proposed method uses multi-scale CNNs to capture features at different scales and attention mechanisms to focus on the most relevant features. The authors evaluate their model on two widely used datasets, the Berlin Database of Emotional Speech and the Ryerson Audio-Visual Database of Emotional Speech and Song, and compare it with several state-of-the-art methods.

The results show that the proposed model outperforms other methods in terms of accuracy and computational efficiency. The authors also provide a detailed analysis of the model's performance, showing that the multi-scale CNN and attention mechanism contribute significantly to its accuracy.

1. **Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining**

The proposed method consists of two phases: text mining and speech mining. In the text mining phase, the authors use natural language processing techniques to extract features from the transcribed text of the audio conversation.

The extracted features are then fed into a support vector machine (SVM) classifier for emotion classification. The authors evaluate their approach to the IEMOCAP dataset, which consists of audio conversations between two actors simulating various emotional scenarios, and compare their results with several state-of-the-art methods..

1. **An effective approach for emotion detection in multimedia text data using sequence-based convolutional neural network**

The authors introduce a novel architecture called Sequence-based Convolutional Neural Network (S-CNN) which employs a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture the temporal and spatial information in text data. The proposed model outperforms several state-of-the-art models on multiple benchmark datasets, demonstrating the effectiveness of the S-CNN architecture for emotion detection in text data.

1. **MULTI-HEAD ATTENTION FOR SPEECH EMOTION RECOGNITION WITH AUXILIARY LEARNING OF GENDER RECOGNITION**

The authors introduce a multi-head attention mechanism to enable the model to focus on different parts of the speech signal and extract relevant features for emotion recognition. Additionally, the model incorporates gender recognition as an auxiliary task, which helps to improve the accuracy of emotion recognition. Experimental results demonstrate that the proposed model outperforms several state-of-the-art models on multiple benchmark datasets, highlighting the effectiveness of the attention mechanism and the auxiliary gender recognition task.

1. **Speech Emotion Recognition Using Speech Feature and Word Embedding**

The authors propose a multimodal fusion model that combines acoustic features extracted from speech signals and text-based features extracted from transcriptions of the speech. The proposed model employs a deep learning architecture incorporating convolutional and recurrent layers. The authors compare the performance of their proposed model with several other fusion models, as well as with models that use only acoustic or text-based features. Experimental results demonstrate that the proposed multimodal fusion model outperforms all other models on multiple benchmark datasets, indicating the effectiveness of the fusion approach for emotion recognition from speech data.

1. **FUSION APPROACHES FOR EMOTION RECOGNITION FROM SPEECH USING ACOUSTIC AND TEXT-BASED FEATURES**

The proposed model combines acoustic features extracted from speech signals and text-based features extracted from transcriptions of the speech. The authors employ a deep learning architecture that incorporates both convolutional and recurrent layers to extract relevant features from the input data. The proposed multimodal fusion model outperforms other fusion models and models that use only acoustic or text-based features, demonstrating the effectiveness of the fusion approach for emotion recognition from speech data.

1. **Audio-Textual Emotion Recognition Based on Improved Neural Networks**

The proposed model incorporates a gated recurrent unit (GRU) with self-attention and a convolutional neural network (CNN) to extract relevant features from the input data. The authors demonstrate the effectiveness of their proposed model on a benchmark dataset and show that it outperforms several state-of-the-art models for emotion recognition from audio and text data.

1. **DEEP MULTIMODAL LEARNING FOR EMOTION RECOGNITION IN SPOKEN LANGUAGE**

The proposed model employs a combination of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to extract relevant features from both acoustic and linguistic data. The authors also incorporate attention mechanisms to enable the model to focus on important parts of the input data. Experimental results demonstrate that the proposed model outperforms several state-of-the-art models for emotion recognition in spoken language.

1. **SPEECH EMOTION RECOGNITION USING MULTI-HOP ATTENTION MECHANISM**

The proposed model employs a combination of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to extract relevant features from the input speech signal. The authors introduce a multi-hop attention mechanism that enables the model to attend to different parts of the input signal at different levels of granularity. Experimental results demonstrate that the proposed model outperforms several state-of-the-art models for speech emotion recognition.

1. **Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network**

The proposed model employs a combination of acoustic and linguistic features and incorporates a multi-level multi-head fusion attention-based recurrent neural network (RNN) to extract relevant features from the input data. Experimental results demonstrate that the proposed model outperforms several state-of-the-art models for speech emotion recognition.

1. **A Bimodal Approach for Speech Emotion Recognition using Audio and Text**

The proposed model employs a convolutional neural network (CNN) and long short-term memory (LSTM) network to extract relevant features from the input data. The authors also incorporate attention mechanisms to enable the model to focus on important parts of the input data. Experimental results demonstrate that the proposed bimodal approach outperforms several state of these techniques.

1. **MULTI-MODAL EMOTION RECOGNITION ON IEMOCAP WITH NEURAL NETWORKS**

The authors proposed a model that combines audio and visual features extracted from the dataset to classify emotions. The audio features are obtained from Mel-frequency cepstral coefficients (MFCCs) and the visual features are extracted from the facial landmarks. The model is based on a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN). The CNN is used to process the visual features, and the RNN is used to process the audio features. The results of the study show that the proposed model outperforms existing methods for multi-modal emotion recognition on the IEMOCAP dataset.

1. **Towards improving e-commerce customer review analysis for sentiment detection**

The authors used a dataset of customer reviews from an e-commerce website and proposed a method to preprocess the data and extract features using natural language processing (NLP) techniques. The authors also proposed a model based on a combination of a convolutional neural network (CNN) and a long short-term memory (LSTM) network to classify the reviews into positive, negative, and neutral categories. The proposed method outperforms existing methods for sentiment analysis on the dataset.

1. **Multimodal Embeddings From Language Models for Emotion Recognition in the Wild**

The authors used a large-scale dataset of online text and speech recordings to train a language model and extract multimodal embeddings. The embeddings are then used as input to a classifier to recognize emotions in speech recordings. The proposed method outperforms existing methods for emotion recognition in the wild, where the data is obtained from various sources with different noise levels and recording qualities.

1. **Speech Emotion Recognition Using Spectrogram & Phoneme Embedding**

The authors proposed a method based on a combination of a convolutional neural network (CNN) and a long short-term memory (LSTM) network to classify the speech into different categories. The proposed method uses universal speech representations obtained from a pre-trained neural network as input. The results of the study show that the proposed method outperforms existing methods for sentiment analysis and emotion recognition on a publicly available dataset.

1. **Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations**

This paper proposes a multimodal approach to recognize emotions in speech signals. The proposed approach utilizes both acoustic features extracted from spectrograms and linguistic features extracted from phoneme embeddings to improve the performance of speech-emotion recognition (SER) systems. The study was conducted on two publicly available datasets, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) and the Toronto emotional speech set (TESS), to evaluate the effectiveness of the proposed approach.

he results show that the proposed approach outperforms the traditional acoustic feature-based SER systems on both datasets. Specifically, the proposed approach achieves an average accuracy of 65.35% and 81.66% on the IEMOCAP and TESS datasets, respectively.

## 3.2 Gap Analysis

All the approaches are not made to have the functionality of extracting emotion through voice and text at the same time. Many of the existing approaches had good architectures but based models were not supposed to be as they wanted to create. Some of the researchers have been using not suitable datasets.

Alternatively, we will create this system and will choose the best accuracy.

Its better accuracy and models still need to be resolved which we will cover through our research and approaches. There will be the use of modern technology and the best models we can use for creating this system.

As a result, our research and model will predict the emotion of humans whether they are happy, sad, angry, love, fear, surprise and neutral. Overall, the previous papers cover different aspects of emotion recognition and sentiment analysis, with some focusing on specific modalities, such as speech or text, and others using a multimodal approach. They also differ in the techniques used for feature extraction and classification, highlighting the diversity of approaches in this field.

While all the papers focus on multimodal emotion recognition, the approaches used are different, and there is no overlap in the techniques used but still, they are not able to find a model where they can depend upon.

# Proposed Methodology

## 4.1 Suggested Approach

We have learned from previous research that emotion recognition is really a complex and difficult task even for humans and for machines and to train them we would have to work hard because it plays a very crucial role in making these machines and technologies user-friendly. Through our research we focus completely on Machine Learning and Deep Learning concepts, using Natural Language Processing (NLP) techniques to develop a system that can predict the emotions of a human. Here we position ourselves in the field of AI-based technologies for emotion recognition, leveraging ML algorithms to detect emotions from written texts and speech data. Our research design involves collecting audio data and converting it into text. We then process and clean this text data using NLP techniques to predict the emotions of individuals, such as happiness, sadness, anger, surprised, fear, love, or neutral. We have read and completely studied every single approach and have gone through its knowledge to create our own system.

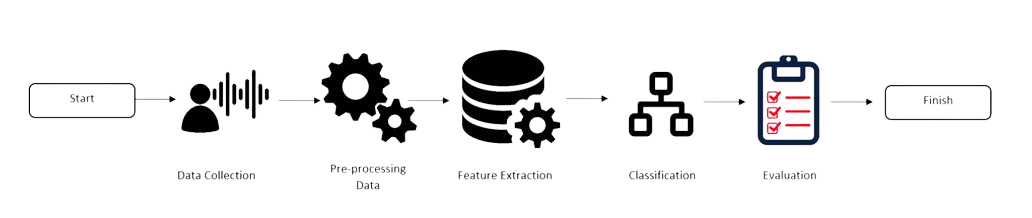


Figure 1 Suggested Approach

We have learned a lot through research by conducting a thorough review of research papers on emotion recognition with different technologies and techniques. From our previous research, we have seen many of the different challenges and what adaptive systems comes to serve it as solutions. We have also seen the advancements being made in emotion recognition systems.

Going through our research design we have to study a lot of research deeply and then we create a combination of computational and quantitative approaches to create an emotion recognition system that can detect at its best. We have followed a basic pattern through starting with the processing and procedures including like data collection, pre-processing, extracting feature selection, training data, model building, and then the evaluation of the whole research. Everything was associated with the effectiveness of our proposed model to detect human emotion.

To develop an appropriate research strategy we have gone through a complete analysis of the problem and its domain with things related to it and also for research objectives completely. We have completely gone through and reviewed every existing technology and approach being made within the strategy to develop such a system. We collected all the resources that we can use whether it’s the technology or tools.

For our project, we have used every source of data we needed. Kaggle.com is the website from where we got our datasets of voice and text also. We have to apply different algorithms to make the datasets useful for us. Where different people's speeches and writings were included which shows the emotion of humans and to which our model is supposed to give the outcome of. Similar to sampling procedures we combine different sampling techniques to ensure the updated approach to be done and for extra diversity. Then we retain the data as 80 percent of it for testing and 20 percent to train it for model development and evaluation.

## 4.2 Workflow of the system

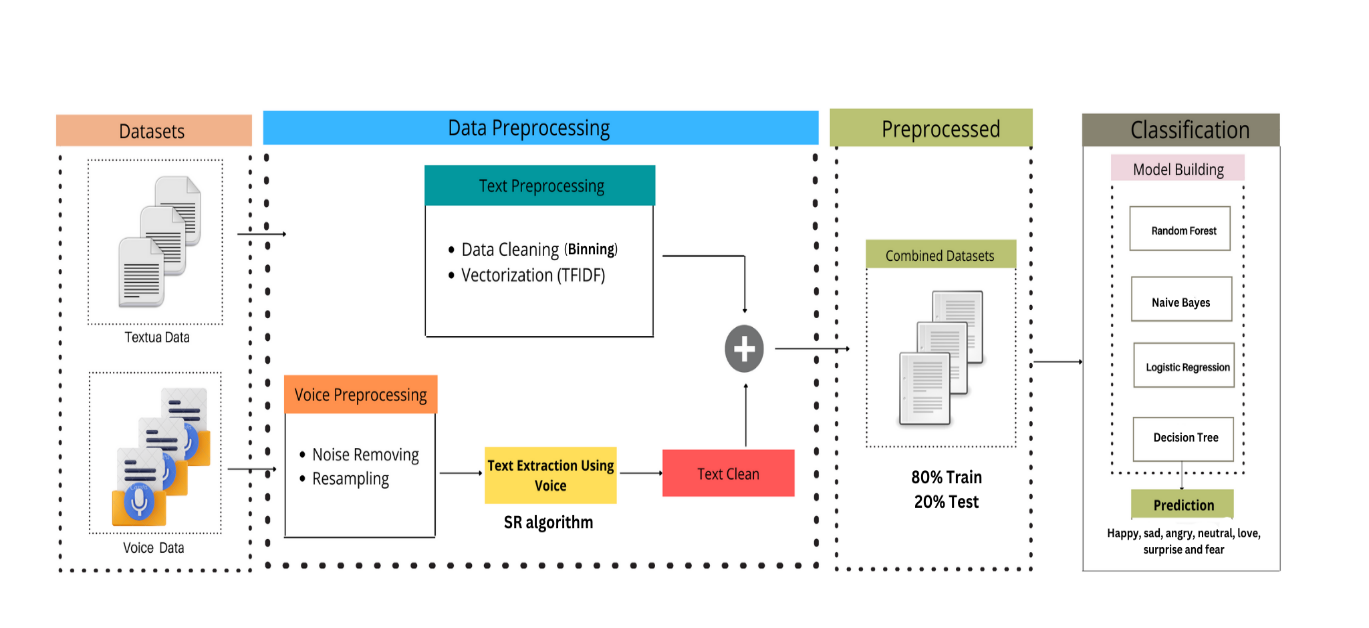


Figure 2 Workflow of the system

To understand the complete workflow of our system firstly we are going through the core setting of our system we have to look at available resources and the nature of the task that how we are going to deal with it. We have chosen techniques involving ML with the NLP for every processing. And towards its core, the main work done is by the programming language (Python) with different libraries and required frameworks for evaluation and implementation.

To understand the core settings of implementation of those tools and techniques deep knowledge of everything was required about ML with the NLP techniques and its toolkit. Python and its libraries with the Tensor Flow, and NLTK with vectorization TFIDF are the main fundamentals to have knowledge about. With a complete understanding and implementation of these techniques, we should have deep knowledge of them.

In the workflow of the system, the calculations involved different processing and analysis of the data with the mean features and emotions. There are different techniques have been used as from converting the audio to text using a transcript model with the algorithms then merging the text files into the final and implementing the text preprocessing techniques which involved data cleaning, lemmatization, vectorization, and after that classification using different models as Random Forest, Logistic Regression, RNN, and LSTM. The procedure involved different processes including data processing, feature selection, data extraction, implementation and model building, and at last the evaluation. As hardware is the basic need to settle down and for that with the different equipment we have used systems that have work with the CPU as well as GPU to handle every situation we face.

As going to the workflow the limitations of the system were to depend on the quality and different responses of the collected data so that accurate prediction can be made of emotion. We assumed that it will show the range of emotions with good accuracy as a result as the latest algorithms and techniques were used for this suitable. For the range of validity, we have to set a scope within the system so it can predict emotions within the language, culture, and influences aspects with a specific content that we needed.

## 4.3 Algorithms/Architecture

We proposed a system using Machine Learning with the NLP techniques which is itself a novel and procedure with a diversity of algorithms. We use a different combination of techniques and tools to the latest technologies we can.

Here we introduce a hybrid or multi-model which can detect human emotion through voice and text at the same time and we also used the different data sets for this purpose. As we used the text datasets and audio then transcribe them to text and implement all the techniques to the text and also data cleaning and different procedures to text. We applied various ML algorithms with NLP techniques with the Random Forest, Logistic Regression, DT, and NB. Then we implemented the system and got the accuracy within the limits and also created a confusion matrix and classification report.

Our research incorporates with different existing systems, as ML and NLP are the major approaches to the system we made. There were existing systems and with multiple different approaches. But we have introduced a system that is hybrid which can detect emotion through voice and text as well at the same time with the latest tools and techniques. We have implemented different algorithms with the classification for NLP processing tasks. We have successfully deployed it to the Logistic Regression, DT, RF, and NB. We have set these things accordingly to gain the maximum output with accuracy and using of modern technologies so we can get the optimal performance as it would be required. In the end, going through this we have developed a system that is now able to detect the emotion of humans with the ML and NLP techniques in manners to predict the human emotion as happy, sad, angry, love, fear, surprise or neutral.

# Design and Implementation

## 5.1 System Design

The performance of the system is designed to achieve the best outcome with an efficient and fast responsive emotion recognition system. Through which humans text and audio speeches can predict emotion in real-time.

The robustness of the system is capable of recognizing the inputs with different parameters such as range of quality, accents, speech patterns, and also support to different writing styles. It should be designed as well so that it can predict the emotion of humans if the voice is not clear have impurities and noisy sounds.

Through the interactivity, the system will be interacting user-friendly as this is the base for it and will provide timely and appropriate responses based on its detection of emotion. It will adapt to the environment and surroundings to which it will predict the best and most accurate emotion.

In flexibility to the system is designed to it will be adaptive to different writings and audios whether speeches or social platforms writings and will provide the emotion according to that whether it’s a post of someone or recording through any means.

In designing to its re-usability and portability system can be integrated with other applications and different systems. It is based on Python and with the latest technologies and advancements so would be accessible by any other machines as technology is in advanced. It is well documented and explained through each phase and its uses.

In security to the system design it ensures the aspects of protecting user data and ensuring data collection, storage, and processing keeping that information just within the system. It is confidential with complete integrity and availability.

## 5.2 System Implementation

1. **Tools and Environment Used:** In this system, code is implemented by using Python programming language with the Jupyter Notebook for which the environment is used by Google Collab. The tools and libraries used include SR, Pandas, and Scikit-learn for speech-to-text conversion, data manipulation, and machine learning algorithms. For storage, google drive is used also to access the dataset.
2. **Implementation of modules:** There have been many of the models used for this system to be designed:

* **NLP techniques:** The complete system is designed and meant to imply the tokenization, lemmatization, and stop words removal which is done by NLTK library. NLP library NLTK has been used widely for these processes. And then the data is pre-processed.
* **Data Preprocessing:** At the very first the audio files have to be converted into text using the Speech Recognition library. It then iterates the labeled folders, processed the files, and organized them into labeled text and also labeled as pandas Data Frame.
* **Text Integration:** It combines the audio-converted audio-to-text dataset with the text datasets of train, test, and value to a final CSV file.
* **Text Preprocessing:** It uses different algorithms to remove the impurities from the text of the final file as removing punctuation and numbers, converting the upper case letter to lowercase, and different preprocessing of data.
* **Train-Test:** Then the dataset is splits into different parts using the train\_train\_split function from sci-kit-learn.
* **Vectorization:** After all the processes the text is finally converted into numerical values of vectors using algorithms of TF-IDF vectorization.
* **Classification:** The code is trained and evaluated with the classification models including, Logistic Regression, Complement Naïve Bayer, Decision Tree, and Random Forest using different vectorization data that is before processed and then test and trained.

1. **Sample Codes and Standards**: The code is integrated with the different systems as designed before and there is a novelty in it and changes have been made. The coding which has been before lacks of the things as detailed explanations and comments which is always a difficult task for everyone to understand and get the concept of it. Through our code, we have set the standards of code clarity and maintainability and make it very easier to understand with the use of the latest technologies in a very simpler way which makes everyone understand it better.
2. **Difficulties faced and addressing:**

* **Datasets:** First we have to combine all the datasets and for that, we have to require the labeled datasets of text as well audio.
* **Audio File:** There have been many formats and as well noise problems that are not clear and hence will not result in a cleared text. However, it supports the format in which most likely to audio will be. The audio file is then meant to be converted into a text file and as accurately as it can be.
* **Text Pre-Processing:** For text pre-processing, we have to get things done efficiently time by time. First, we removed punctuations, then upper case letters and sentence making so it required modification every time we go through it.
* **New Technologies:** Working on new technologies was a fundamental and crucial challenge for us. As it has to be working perfectly for the output of the system. We have to research and get each tool and technology to work perfectly.
* **Runtime Efficiency:** It has to be done in a way that runtime efficiency has to be corrected and more frequently in use of the system we have designed for our system. And everything depends upon the runtime.
* **Model building:** Model building requires the perfect knowledge and research of it and very careful optimization to achieve the best performance.

## 5.3 Assumptions/Constraints (Optional)

* Audio Dataset: The audio dataset is assumed that it has good audio files.
* Labeling Accuracy: The dataset must be labeled well which we have already checked.

# 6. Evaluation

## 6.1 Experimentation

Our project outlines the experimental setup, design, and each detail that can obtain the required results to detect the emotion. For that purpose, actual work is done as:

1. **Installation of Libraries:** First of all we have to install the libraries here as Speech Recognition and Pandas using pip instructions. These are the necessary libraries.
2. **Mount Google Drive:** For using the dataset we have to mount the google drive so it can be used flexibly where we want it to be.
3. **Data Collection:** It has to collect the over data from datasets of audio files and text files within the CSV file. It creates a loop where we start converting audio into text one by one till the last audio. Then the text and labeled data as CSV file was stored in pandas.
4. **Combining Datasets:** Then we combined the audio-converted text file dataset with already datasets we are having of people's writings. We combined all the files of text datasets as a final CSV to do operations on it.
5. **Text Preprocessing:** We have done text pre-processing where it has removed any of the punctuations, numbers, and single characters. We have also implemented the conversion of upper case letters into lower ones so everything will be well organized into a CSV file.
6. **Train:** The labeled dataset of the CSV file then is trained through train\_test\_split and learn.model\_selection module. This is done for the training and evaluation of machine learning models.
7. **Text Vectorization:** Then we have to convert the preprocessed text data to a numerical value by vectorization for reaching out maximum output with the TF-IDF vectorizer. Then these vectors are representing each word present in the text data.
8. **ML models training and evaluation**: The code is trained and evaluated with different machine learning models including Logistic Regression, Decision Trees, Complement Naïve Bayes, and Random Forest Classifier. As the trained data is used for these models to find the accuracy.
9. **Confusion Matrix:** To get the predictive and actual values of these models we have created the confusion matrix with the classification report.

### 6.2 Experimental Setup

Experimental setup involves the project tools, environment, and resources being used. We have used:

* Tools and Libraries: In our code, we have used several tools and libraries such as Speech Recognition, Pandas, and sci-kit-learn.
* Environment: The environment of our project is doing code on Collaboratory Notebook, which is an online platform for Python Programming specifically.
* Data Path: The complete project emphasizes the use of audio and text datasets that were in the google drive as the path we mounted the code within the drive.
* Data Pre Processing: In Data Pre Processing we have converted audio files into text using Speech Recognition techniques.

### 6.3 Experiments Design/Details

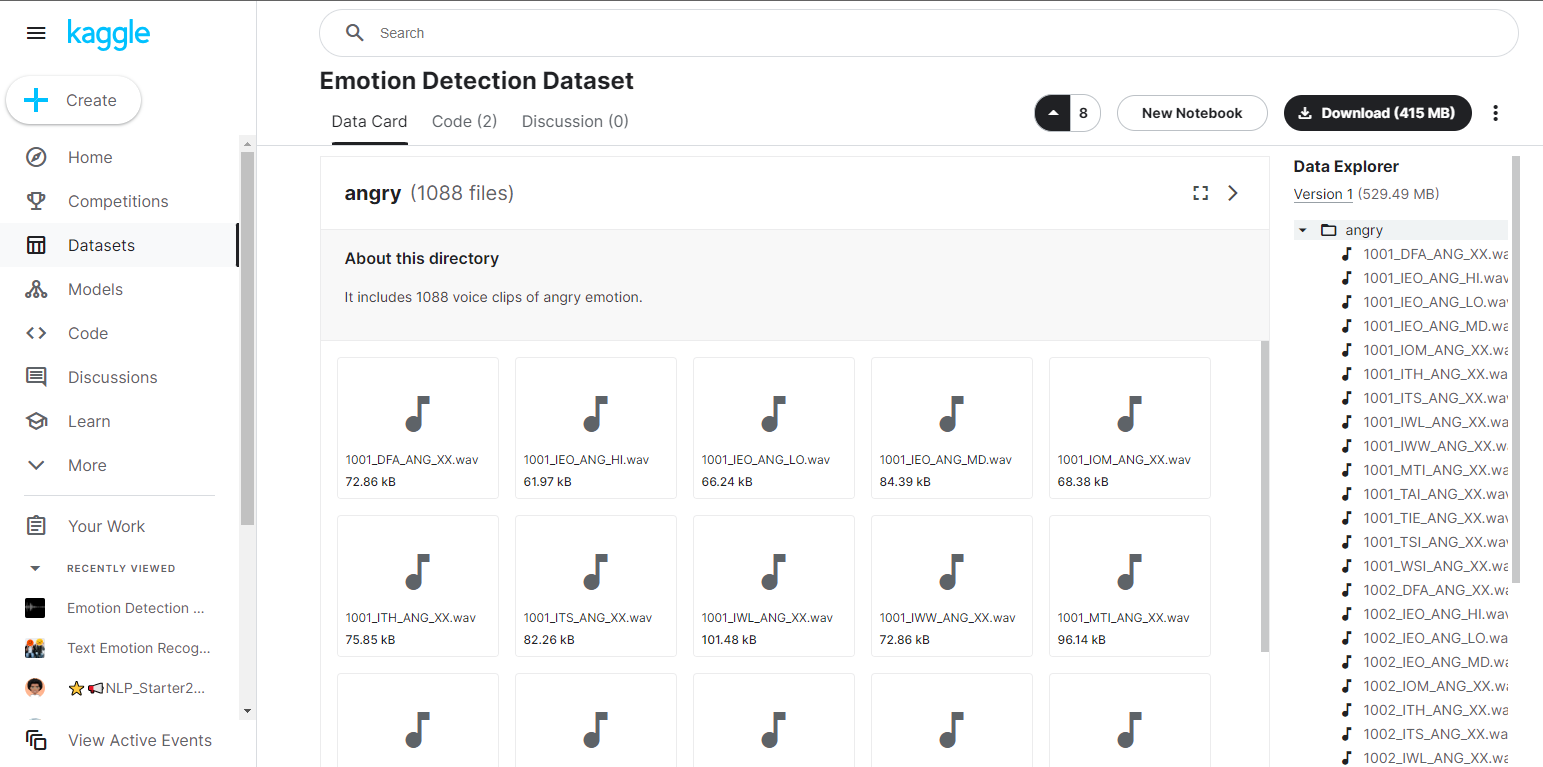
Experiment design in this project was meant to provide a system that can detect the emotions of humans. For that we have used:

* **Data Collection:** In this project, we have first collected the data and then we implemented the conversion of audio files into text using a loop we were able to convert all the existing audio files using Speech Recognition.
* **Data Frame Creation:** We created an empty data frame to store the text and associated labels for the dataset. As data frame has two columns one is text and the other is the label.
* **Text Pre-Processing:** Then the converted data has to go with the pre-processing period in which we combined all the datasets into a final CSV and implemented the techniques to remove punctuations, numbers, and single characters.
* **Data Label Standardization:** Then we labeled the dataset we created through audio files and text dataset by replacing labels as our model requirement. ‘Joy’ with the ‘happy’ ‘sadness’ with ‘sad’ and ‘anger’ with ‘angry’ as we are predicting emotion of humans it has to be properly labeled as emotion.
* **Training and testing:** Then the dataset of the final CSV has been trained to different models after testing and then for validation.
* **Machine Learning Models:** We have implies different machine learning models such as Logistic Regression, Decision Trees, Complement Naïve Bayes, and Random Forest Classifier, and then it evaluates the performance by confusion matrix and classification report. Results

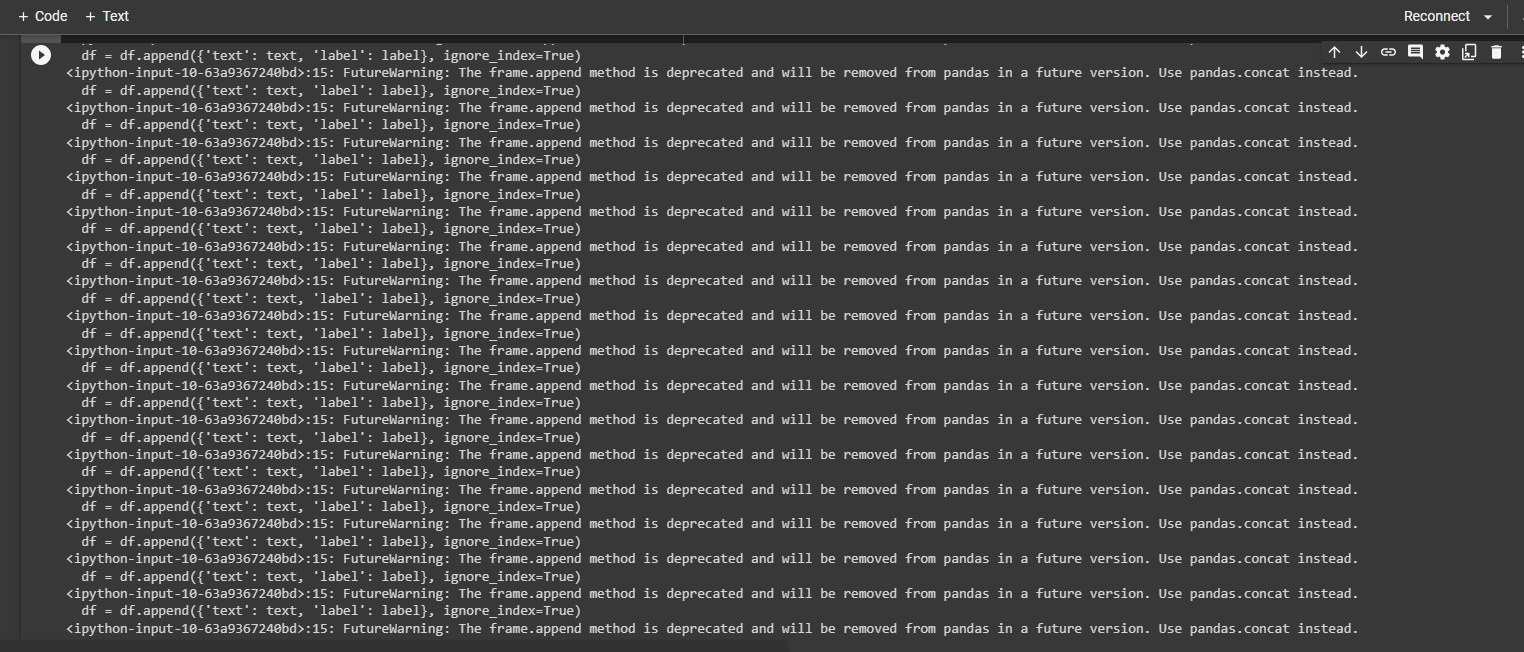
## Discussion/Analysis

## 6.4.1 Dataset:

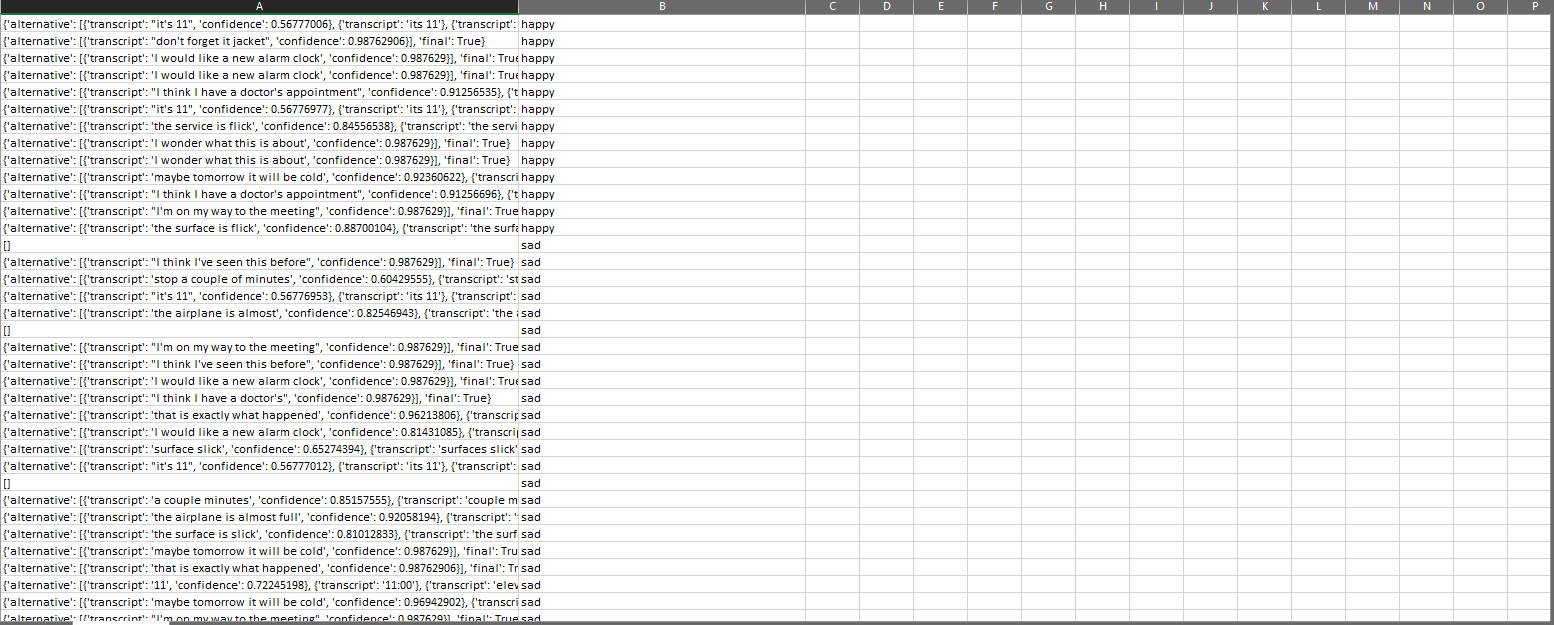
* **Audio Dataset:** We used the dataset from kaggle.com which has the happy, sad, angry, and neutral audios in it. The audios are in wav format which is easily accessible and easy to work on.



*Screenshot 1 – Audio Dataset*

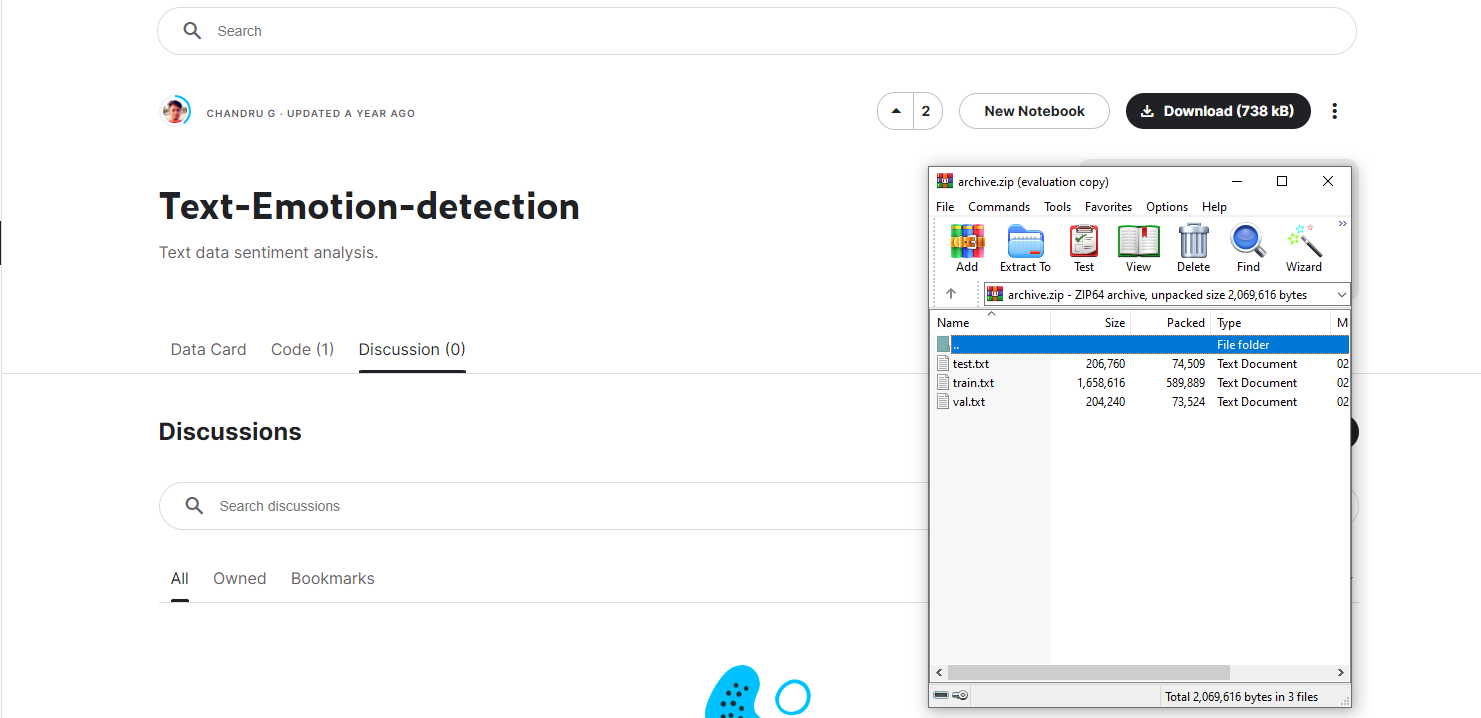


*Screenshot 2 - Audio Dataset Conversion*



*Screenshot 3 - Converted CSV File*

* **Text Dataset:** We use the text datasets from kaggl.com which were labeled data within the txt file which we converted into CSV file for our system.

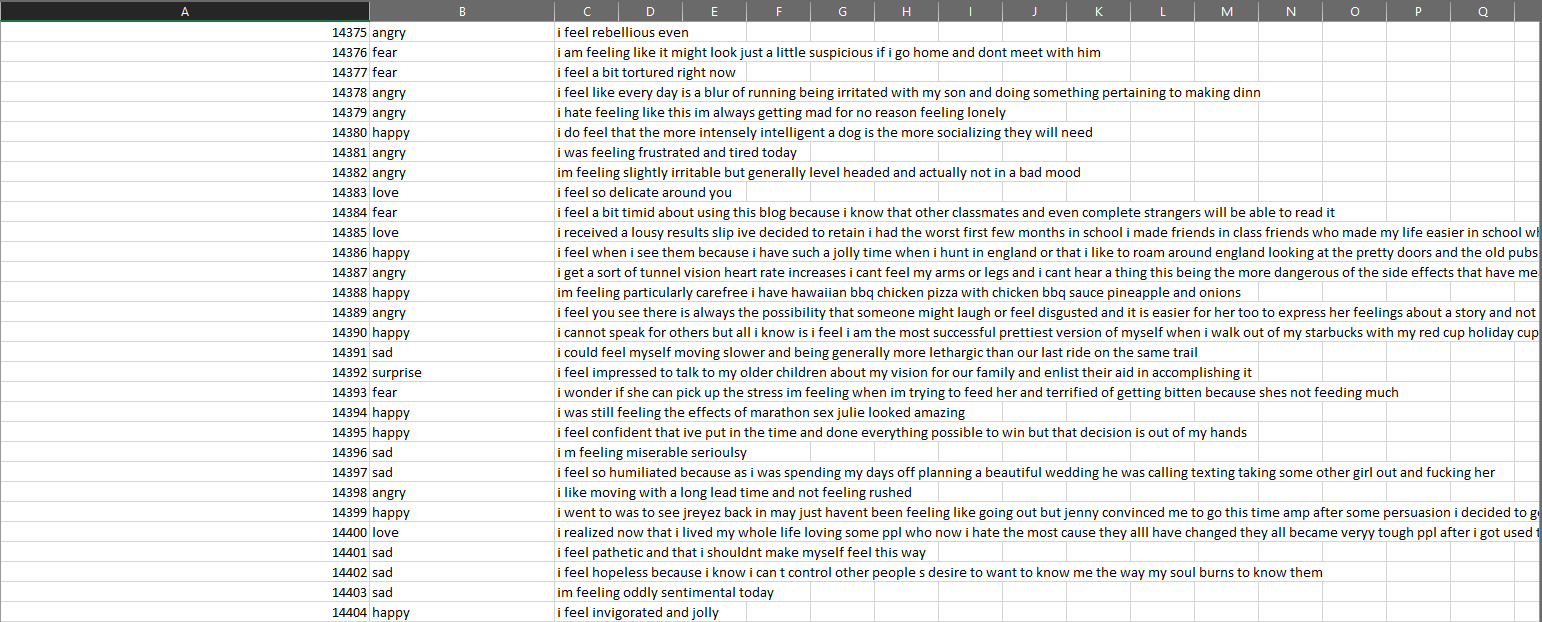


*Screenshot 4 - Text Dataset*

* **Combined Dataset:** Then we combine the converted audio to text dataset with the already text datasets of value, train, and test. We combine these all into a CSV file.



*Screenshot 5 - Combined Dataset*



*Screenshot 6 - Combined Dataset CSV*

## Models:

We evaluate the system into four models as Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest.

## Logistic Regression:

Logistic regression model is trained on a set of vectors with train vectors and with different and corresponding and target labels y\_train. Then the model was tested again on test\_vectors and also to correspond to target labels y\_test.

The logistic regression model is trained to a set of vectors and corresponding target values to y\_train and then the model was tested on vectors test\_vectors and exact to its labels as y\_test.

The confusion matrix is the representation of model performance. As each row is for the true classes, and each column is for the predicted classes. Values within the whole matrix show the number of instances that were classified accordingly. As this model predicted 522 instances as angry, 374 instances as fear, and 1276 as happy, and so on by the evaluation of this model. Hence the equation for Linear Regression as for Machine Learning models will be

By getting into each value of this we will be containing:

After this if we get between the ranges between –[infinity] to +[ infinity] then the equation will be based on the log with:

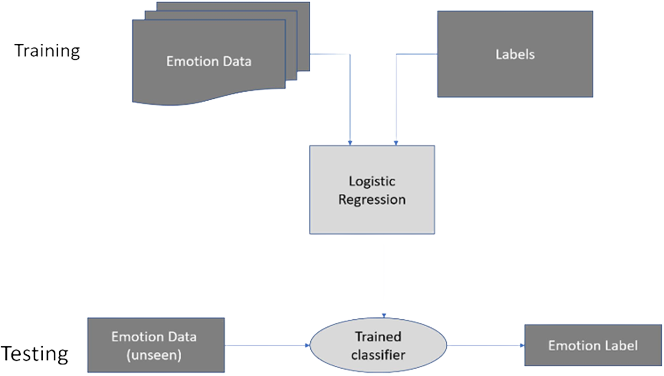
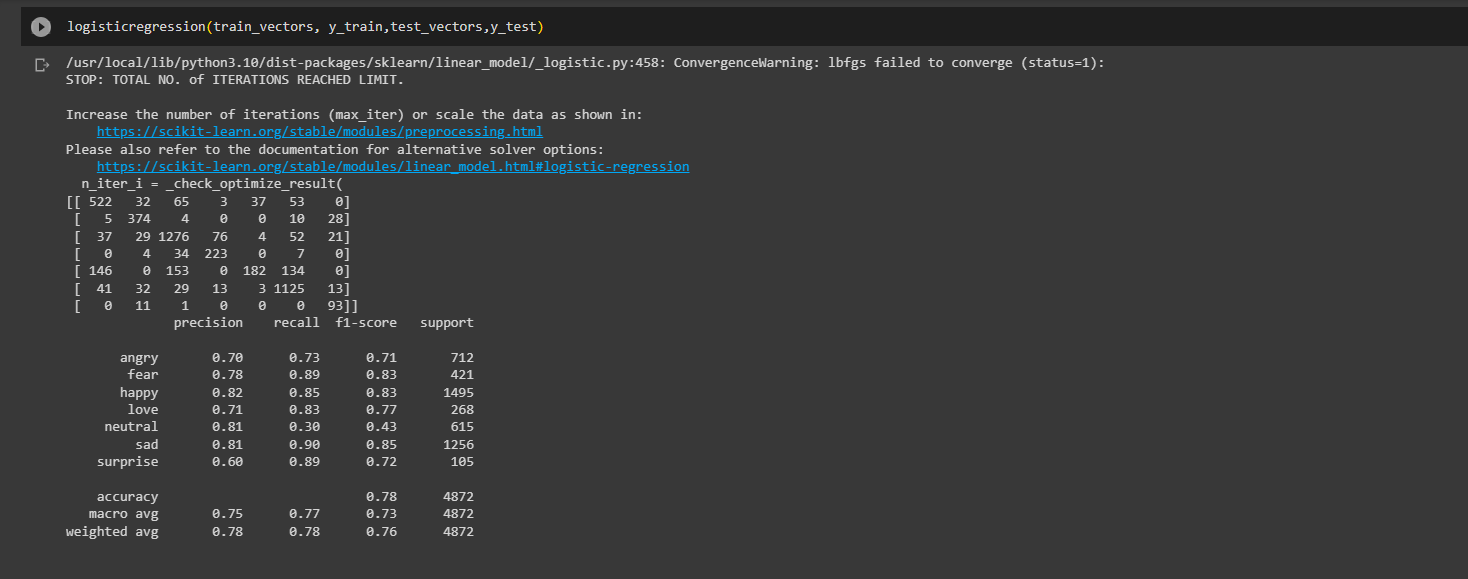


Figure 3 Logistic Regression



*Screenshot 7 - Logistic Regression*

Table 4 Confusion Matrix LR

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 522 | 32 | 65 | 3 | 37 | 53 | 0 |
| 5 | 374 | 4 | 0 | 0 | 10 | 28 |
| 37 | 29 | 1276 | 76 | 4 | 52 | 21 |
| 0 | 4 | 34 | 223 | 0 | 7 | 0 |
| 146 | 0 | 153 | 0 | 182 | 134 | 13 |
| 41 | 32 | 29 | 13 | 3 | 1125 | 13 |
| 0 | 11 | 1 | 0 | 0 | 0 | 93 |

## 6.4.2.2 Naïve Bayes:

The system with using a model of Complement Naïve Bayes (CNB) classifier implements in sci-kit-learn to perform the text classification of this model. The model here is calling trains to the CNB model and data is tested and trained on it with the predictions.

As a result of that confusion, the matrix is providing a summary of model predictions as different labels across it. It is displaying the positive, and negative values respectively. Through the confusion matrix we can analyze the correct and incorrect predictions based on that result will be also. While classification report of this model provides the evaluation of performance in class. It includes different matrices such as recall, f1-score, and different support. It measures the accuracy of positive predictions and as well calculates the proportion of actual positive identification of the model. These matrices help us to evaluate the effectiveness of the Naïve Bayes models as classified in text and provided data. The general equation for Naïve Bayes in ML is

While condition the probability can also detect each class problem. And to provide that Naïve Bayen is using the calculation simple formulae as

As for the model to detect emotion here will be as generalized:

**Here P(C|X):** Probability of a class given with x features.

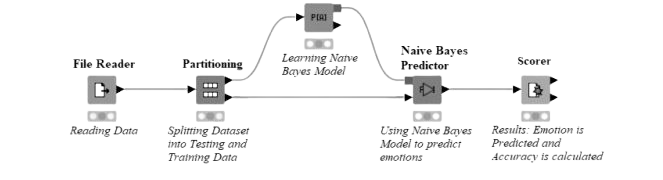
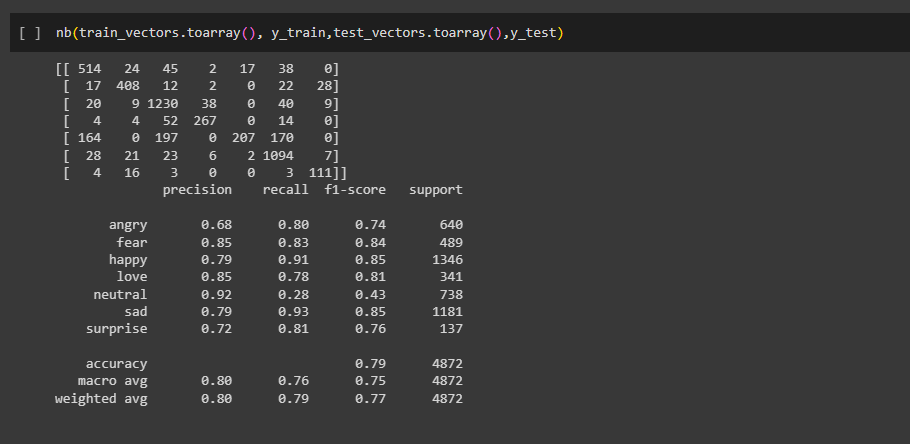


Figure 4 Naive Bayes



*Screenshot 8 - Naive Bayes*

Table 5 Confusion Matrix NB

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 514 | 24 | 45 | 2 | 17 | 38 | 0 |
| 17 | 408 | 12 | 2 | 0 | 22 | 28 |
| 20 | 9 | 1230 | 38 | 0 | 40 | 9 |
| 4 | 4 | 197 | 0 | 207 | 170 | 0 |
| 164 | 0 | 197 | 0 | 207 | 170 | 0 |
| 28 | 21 | 23 | 6 | 2 | 1094 | 7 |
| 4 | 16 | 3 | 0 | 0 | 3 | 111 |

## 6.4.2.3 Decision Tree:

Through the decision model we determine the features of predition it is giving. For this there are two methods as Gini Index and Entropy. Entropy is used to measure information content in a present node while Gini Index is used for measuring the impurity of selected node.

Gini Inde: Gini Index for given node as t will be:

Where we know that is the emotion level i as in node t.

Entropy:

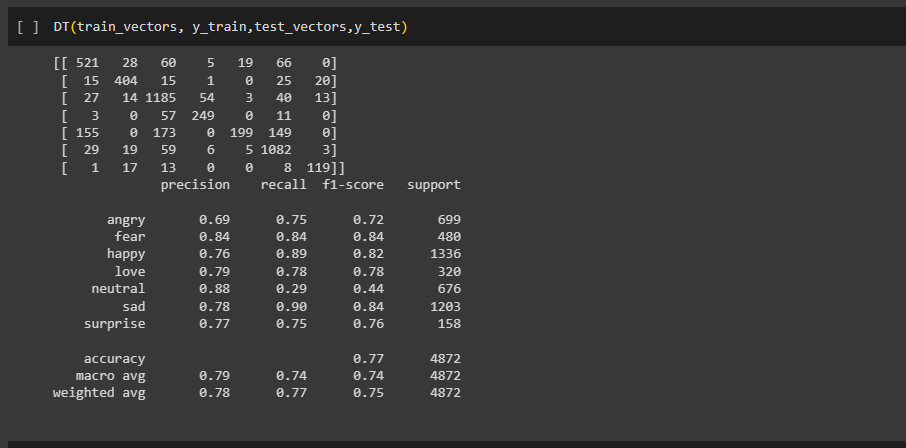
For a given node t as entropy will be:

Where we know that is the emotion probability for any label i as in node t.

This both functioning of Gini Index and Entropy will gives us the best feature of model.



Figure 5 Decision Tree



*Screenshot 9 - Decision Tree*

Table 6 Confusion Matrix DT

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 521 | 28 | 60 | 5 | 19 | 19 | 66 | 0 |
| 15 | 404 | 15 | 1 | 0 | 0 | 25 | 20 |
| 27 | 14 | 1185 | 54 | 3 | 3 | 40 | 13 |
| 3 | 0 | 52 | 249 | 0 | 0 | 11 | 0 |
| 155 | 0 | 173 | 0 | 199 | 199 | 149 | 0 |
| 29 | 19 | 59 | 6 | 5 | 5 | 1082 | 3 |
| 1 | 17 | 13 | 0 | 0 | 0 | 8 | 119 |

## 6.4.2.4 Random Forest:

Bootstrap Aggregating: Here randomly selecting any bootstrap sample from the size N:

A calculated subset for the model:

Now the model will start the Decision Tree Training with T1 on a bootstrap with sample D1 and using the feature as F1 so after that doing the Voting and Aggregation model will be repeating the steps 1-3 times for decision tree as:

Hence from the following instances, it will make predictions using decision tree as:

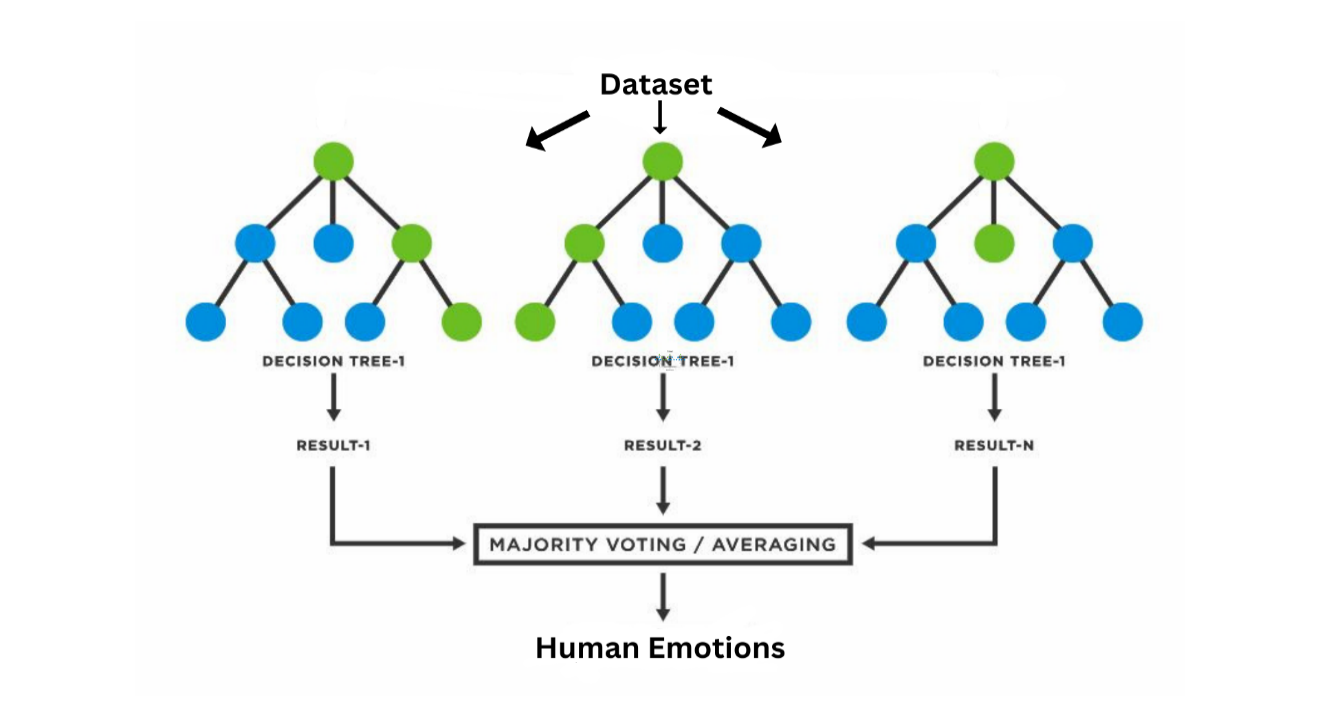
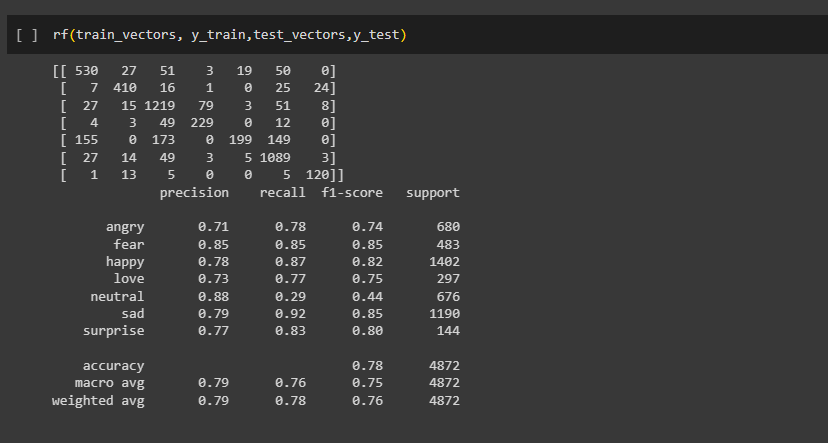


Figure 6 Random Forest



*Screenshot 10 - Random Forest*

*Table 7 - Confusion Matrix RF*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 530 | 27 | 51 | 3 | 19 | 50 | 0 |
| 7 | 410 | 16 | 1 | 0 | 25 | 24 |
| 27 | 15 | 1219 | 79 | 3 | 51 | 8 |
| 4 | 3 | 49 | 229 | 0 | 12 | 0 |
| 155 | 0 | 173 | 0 | 199 | 1089 | 3 |
| 27 | 14 | 49 | 3 | 5 | 1089 | 3 |
| 1 | 13 | 5 | 0 | 0 | 5 | 120 |

## 6.4.3 Results:

The results obtained from the implementation of code defines the effectiveness of speech recognition of converting from audio to text and then the accuracy for the emotion recognition system is classifying the emotions. The results provide a very fine and well accurate system to detect the human emotion. The obtained results through this project align with the expectations and hypothesis set.

## Evaluation:

The implementation successfully explored the relationship between the speech patterns and emotional states of humans as emotions are really complex but everything working properly from start to end makes it easier to detect.

## 6.4.5 Comparative Analysis:

The results of this system is consistent and well accurate from the previous research. Other research that has been made through and creating systems like these faced many challenges in detecting and distinguishing the emotion of humans and to which our model worked properly. The findings validation of our model gives us the effectiveness of speech-based approaches.

## Different Researches and their proposed models with methodology and results:

Table 7 Comparative Analysis Research Papers

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **Problem** | **Methodology** | **Model** | **Results** |
| Emotion detection from text and speech | GMMs, SVM, HMM, Neural Network | CNN joint learning | 85% Accuracy |
| Speech Emotion Recognition Using Deep Learning Techniques | DL and SER | RNN | 61 % Accuracy |
| state-of-the-art approaches for emotion recognition in text | DL, Hybrid | SVM | 67 % Accuracy |
| Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep Bi-LSTM | SER | CNN, SER | 72 % Accuracy |
| Speech Emotion Recognition with Multi-task Learning | GMMs, SVM, HMM, MLP | DNN-GMM | 83.97 % Accuracy |
| Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining | Hybrid, DL | UBM,GMM | 69.2 % Accuracy |
| Deep learning approach to text analysis for human emotion detection from big data | Hybrid, SVM, NLP | SVM | 90 % Accuracy |
| SPEECH EMOTION RECOGNITION USING SELF-SUPERVISED FEATURES | NLP,DL | DLSTA | 97.22 % Accuracy |
| Multimodal\_Speech\_Emotion\_Recognition\_Using\_Audio\_and\_Text | NLP techniques | Upstream + Downstream | 77.76 % Accuracy |
| Multi-model Speech emotion recognition using audio and text | DNN, RNN | ASR, TRE | 43.51 % Accuracy |
| Detecting Emotion from Natural Language text using hybrid and NLP models | NLP, ML,DL | Bi-LSTM, Hybrid | 61 % Accuracy |
| Multi-model emotion recognition with high level speech and text features | CNN, Word2vec and GloVe | CNN, SER,TER | 70.1 % Accuracy |
| Bimodal speech emotion recognition using pre trained language models | LSTM, SER, ASR | SER based XLNET-Base | 69.5 % Accuracy |
| Deep neural networks for emotion recognition combining audio and transcripts | LSTM | LSTM, GRU networks | 68.8 % Accuracy |
| Efficient speech emotion recognition using multi scale CNN and attention | Bi-RNN, MSCNN, ASR | MSCNN-SPU | 74 % Accuracy |

* + - 1. **Our proposed model:**

Table 8 Proposed Model

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem** | **Methodology** | **Model** | **Results** |
| Emotion Recognition through voice and text using NLP techniques | Speech Recognition, TF-IDF | CNB,DT,LR,RF | 80% Accuracy |

## Limitations and challenges:

The accuracy of our system is directly connected to quality and diversity of the training data that we used. There were challenges of noisy audio recordings and unclear sounds so through which it makes difficult for us to measure the quality we want. We have used different algorithms and the latest technologies to overcome this issue.

## Insights:

Our system is a speech-based emotion recognition system that can be widely used as well in applications such as voice assistants, customer sentiment analysis, and mental health monitoring. The accurate analysis of our system makes us a personalized human-computer interaction system which is the main need. This will also enhance the emotion detection accuracy for expressing

# 7. Conclusion and Future work

This project has been successful in detecting the human emotions in different 7 categories as happy, sad, angry, neutral, love, surprise, and fear. We have detected the emotions using audio as well as the text of the people. By using the machine learning algorithms we have successfully deployed the four models as logistic regression, naïve bays, decision tree, and random forest. We obtained the highest accuracy for the naïve bays system. The classifiers we have used gives us the best accuracy with their good performance matrices as accuracy, precision, recall, and F1-score and it detects the emotion of humans well accurately as expected.

The implications of this system are very significant in various domains such as E-commerce, medical, health sectors, robotics, Education sectors, security, social media, business, and elections. Emotion recognition can give value able sights insights for businesses for improving their product, services, and customer experience with a more friendly nature. Where understanding and responding to human emotions are very complex and difficult tasks but through this module, we will be able to detect human emotions accurately.

Detecting human emotions through their speeches and writings is really a complex task that always requires proper and careful consideration of many factors on which our system relies. Meanwhile creating this system our classifiers have shown desired and promising results that are improved and introduced. Future work and results in this area of the field we worked on will expand the knowledge base and can also enhance the performance of the system. We have implemented and expanded our knowledge with different new tools and technologies by its implementation and for future work:

1. **Advance system:**

We have successfully created an advanced AI approaches system that can detect human emotion which is really a complex task. We have used advanced techniques such as word embedding, contextual embedding, and ML models to capture the desired results. These techniques have improved the system on the base of previous systems were made.

1. **Real-time application: In work to the future it can be implemented on a re**al-time application based through which we can record people's speeches and we can detect their emotions on the basis of their speeches and the text they have written. This enhancement will allow different platforms to directly detect human emotions.
2. **Data Imbalance:** In the future data imbalance issues can also be raised as repeating the data to the systems where we will consider to not oversampling, or under-sampling of the data distribution and will also improve the model's accuracy through classifying it to relevant models.
3. **Error Analysis:** There would be different patterns as technology advances and will also enhance or update the systems requirements as well the errors inside. To which overcome the errors that will be faced by classifiers and different models we will also train the model to deal with them.

By addressing these recommendations and conducting more relevant searches on this system we would be able to update the system as well as the requirements that we need to retain it to work effectively and efficiently. The knowledge in this field of areas is very vast and can be expanded by different research and implementations which will then result in the best accuracy given systems. This will always facilitate the better understanding and interaction between humans and machines to work more friendly and understanding of each other. Different advancements in this system will make the accuracy better, more context-aware, and intelligent systems that can facilitate humans better.

# 8. References/ Bibliography

[1] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," IEEE Access, vol. 8, pp. 79861-79875, 2020. [Online].

Available: <https://ieeexplore.ieee.org/abstract/document/9078789>

[2] X. Cai, et al., "Speech Emotion Recognition with Multi-Task Learning," in Interspeech, vol. 2021, 2021. [Online].

Available: <https://openreview.net/pdf?id=K90Yn2bZZ37H>

[3] J. Bhaskar, K. Sruthi, and P. Nedungadi, "Hybrid approach for emotion classification of audio conversation based on text and speech mining," Procedia Computer Science, vol. 46, pp.635-643,2015.[Online].

Available: <https://www.sciencedirect.com/science/article/pii/S1877050915001763>

[4] J. Guo, "Deep learning approach to text analysis for human emotion detection from big data," Journal of Intelligent Systems, vol. 31, no. 1, pp. 113-126, 2022. [Online].

Available: <https://www.degruyter.com/document/doi/10.1515/jisys-2022-0001/html>

[5] E. Morais, et al., "Speech emotion recognition using self-supervised features," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2022.[Online].

Available: <https://ieeexplore.ieee.org/abstract/document/9747870/>

[6] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in 2018 IEEE Spoken Language Technology Workshop (SLT), 2018. [Online].

Available: <https://ieeexplore.ieee.org/abstract/document/8639583/>

[7] S. Madhuri, "Detecting emotion from natural language text using hybrid and NLP pre-trained models," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 10, pp. 4095-4103, 2021. [Online].

Available: <https://www.turcomat.org/index.php/turkbilmat/article/view/5122>

[8] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features," in 2021 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU),2021,pp.84-91.[Online].

Available: <https://ieeexplore.ieee.org/abstract/document/9688036/>

[9] V. Heusser, et al., "Bimodal speech emotion recognition using pre-trained language models," arXiv preprint arXiv:1912.02610, 2019. [Online].

Available: <https://arxiv.org/abs/1912.02610>

[9] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 629-635. [Online].

Available: <https://ieeexplore.ieee.org/abstract/document/8639583/>

[10] Z. Peng, et al., "Efficient speech emotion recognition using multi-scale CNN and attention," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6389-6393. [Online].

Available: <https://ieeexplore.ieee.org/abstract/document/9414286/>

[11] K. Shrivastava, S. Kumar, and D. K. Jain, "An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network," Multimedia Tools and Applications, vol. 78, no. 20, pp. 29607-29639, 2019. [Online].

Available: <https://link.springer.com/article/10.1007/s11042-019-07813-9>

[12] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7794-7798. [Online].

Available: <https://ieeexplore.ieee.org/abstract/document/9054073/>

[13] S. Sailunaz, K. Kashfia, et al., "Emotion detection from text and speech: a survey," Social Network Analysis and Mining, vol. 8, pp. 1-26, 2018. [Online].

Available: https://link.springer.com/article/10.1007/s13278-018-0505-2

[14] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical Signal Processing and Control, vol. 47, pp. 312-323, 2019. [Online].

Available: https://www.sciencedirect.com/science/article/abs/pii/S1746809418302337

# 9. Appendix

## 9.1 Glossary of terms

Table 9 - Glossary of terms

|  |  |
| --- | --- |
| ML | Machine learning is the use and development of supervised systems which is able to adapt and learn new things |
| NLP | Natural Language Processing is the branch of Ai that makes computer comprehend, generate and manipulate human language |
| SR | Speech Recognition is the ability of the system to identified spoken content and convert the audio speeches into text |
| LR | Logistic Regression model of ML is used to predict categorical dependent variables |
| RF | Random Forest model of ML makes decision tree and predict according to majority voting |
| NB | Naïve Bayes model of ML used for quick and constant predictions |
| Dataset | Collection of database |
| DT | Decision Tree is the model implementation based on supervised learning which predicts the answer based on the history |
| TFIDF | Term-frequency- Inverse Frequency Document is to show how relevant is a word to document in a collection of documents |
| NLTK | Natural Language Tool Kit used for building programs that works with the human language data |

## Pre-requisites

**1. Linguistic and NLP fundamentals:** Emotion recognition using natural language processing (NLP) techniques applied to voice and text data is an emerging field of research. To ensure a comprehensive understanding and successful replication of studies in this area, it is crucial to outline the prerequisites or requirements. This section provides an overview of the necessary knowledge, skills, resources, and conditions that are essential for comprehending and conducting research on emotion recognition using NLP through voice and text.

**2. Linguistic and NLP:** To develop the system with NLP, there has to be proper consideration and understanding of linguistics and NLP. We have to consider the knowledge, skills, conditions, and resources that would be very important for conducting research on emotion by using NLP for voice and text at the same time.

**3. Machine Learning and NLP techniques:** Being proficient with machine learning and NLP techniques are very essential for implementing the emotion recognition system. Researchers should have knowledge about algorithms such as Speech Recognition, Decision trees, Logistic Regression, Random Forest, and Naïve Bays with their feature extraction and model evaluation.

**4. Audio signal Processing:** As the system is based on audio signals where people's speeches will be used to detect the emotion, an understanding of signal processing is necessary.

**5. Text Processing and Sentiment Analysis:** Having expertise in text processing techniques will be helpful for using in as we will use the dataset of text data and text processing is the main fundamental process of this system. We would have to get the knowledge of lemmatization, vectorization, and tokenization.

**6. Emotion Theories:** We have to get knowledge of emotion expression completely so we can differentiate between the expression and we can then classify emotion accurately.

**7. Datasets and Annotation:** We have to get the labeled text and voice dataset. As the dataset has a lot of entities we can’t figure out easily what emotion is associated with any writing or speech so have to get the labeled text and speech data.

**8. Previous Research and Literature:** We have completely studied around 30 of the research paper and we got knowledge about things that would have been a must for creating this system. We would need to evaluate systems as unique, novel, and best with accuracy. We gain knowledge of methodologies, studies, and advancements with those researches that have been made earlier.

## 9.3 Reference/ Source Documents

* Research Papers
* Books
* Kaggle.com
* Website and Online Researches
* Online Database