TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING

Khwopa College Of Engineering
Libali, Bhaktapur
Department of Computer Engineering

A PROPOSAL ON

MULTIMODAL EMOTION RECOGNITION FROM TEXT AND FACIAL EXPRESSION

Submitted in partial fulfillment of the requirements for the degree

BACHELOR OF COMPUTER ENGINEERING

Submitted by

Saurab Khatiwoda	KCE077BCT034
Suman Adhikari	KCE077BCT038
Suraj Timilsina	KCE077BCT039
Utshab Timalsina	KCE077BCT047

Under the Supervision of
Er. Sushil Dyopala
Department Of Computer Engineering

Khwopa College Of Engineering
Libali, Bhaktapur
2023-24

CERTIFICATE OF APPROVAL

This is to certify that this minor project work entitled **"Multimodal Emotion Recognition from Text and Facial Expression"** submitted by Saurab Khatiwoda (KCE077BCT034), Suman Adhikari (KCE077BCT038), Suraj Timilsina (KCE077BCT039) and Utshab Timalsina (KCE077BCT047) has been examined and accepted as the partial fulfillment of the requirements for the degree of Bachelor in Computer Engineering.

.....
Er. Ashok GM
External Examiner
Associate Professor
Department of Electronics and Computer Engineering
Himalaya College of Engineering

.....
Er. Sushil Dyopala
Project Supervisor
Machine Learning Engineer
Fuse Machine Nepal Pvt. Ltd.

.....
Er. Dinesh Gothe
Head of Department,
Department of Computer Engineering
Khwopa College of Engineering

Copyright

Acknowledgement

We take this opportunity to express our deepest and sincere gratitude to our supervisor Er. Sushil Dyopala, for his insightful advice, motivating suggestions, invaluable guidance, help, and support in the successful completion of this project, and also for his constant encouragement and advice throughout our Bachelor's program.

Additionally, we would like to thank our HoD Er. Dinesh Gothe for providing valuable suggestions and for supporting the project.

Saurab Khatiwada	KCE077BCT034
Suman Adhikari	KCE077BCT038
Suraj Timilsina	KCE077BCT039
Utshab Timalsina	KCE077BCT047

Abstract

The proposed research, titled "Multimodal Emotion Recognition from Text and Facial Expression", aims to develop a comprehensive approach to emotion recognition by integrating facial expression analysis and textual content processing. Emotions are complex and nuanced, often expressed through both facial cues and linguistic patterns. This study seeks to leverage machine learning techniques to synergistically analyze multimodal data, combining facial features extracted from images with textual information. The integration of these modalities is expected to enhance the accuracy and robustness of emotion recognition models. The research will involve the collection of a diverse dataset, the development of advanced machine learning algorithms, and the evaluation of the proposed approach's performance. The outcomes of this study have the potential to contribute significantly to applications in human-computer interaction, affective computing, and other fields where accurate emotion recognition plays a crucial role. The proposed multimodal emotion recognition system achieves high F1 scores of 0.6658 and 0.9561 for the Custom ConvNet and Fined Tuned BERT models, respectively, with corresponding accuracies of 68.07% and 95.56%.

Keywords: *Multimodal Emotion Recognition, Facial Expression Analysis, Text-based Emotion Recognition, Emotion Classification, Natural Language Processing (NLP), Multimodal Fusion Techniques*

Contents

List of Tables

List of Figures

List of Symbols and Abbreviation

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
ConvNet	Convolutional Network
FER	Facial Emotion Recognition
MELD	Multimodal EmotionLines Dataset
MER	Multimodel Emotion Recognition
NLTK	Natural Language Toolkit
NLP	Natural Language Processing
NumPy	Numerical Python
ROC	Receiver Operating Characteristic
SDLC	Software Development Life Cycle

Chapter 1

Introduction

1.1 Background

Novel applications have been made possible by the convergence of machine learning, computer vision, and natural language processing, particularly in the field of affective computing. In human-computer interaction, the comprehension and interpretation of human emotions are crucial, and combining text-based emotion detection with facial expression analysis appears to be a promising approach. Understanding that emotions are complex and frequently conveyed through written language as well as facial clues, this research tackles the difficulty of creating a Multimodal Emotion Recognition system. By utilizing advanced machine learning algorithms, such as deep learning models, the research seeks to concurrently extract significant insights from textual data and facial expressions. This project has the potential to greatly advance the field of emotionally intelligent AI systems, leading to more responsive and sympathetic human-machine interactions in a variety of settings.

1.2 Problem Statement

The fundamental issue with Multimodal Emotion Recognition utilizing Facial Expression and Text through Machine Learning is the complexity of human emotion expression, which is frequently complicated and communicated through a variety of media, including written communication and facial expressions. While current emotion detection algorithms are capable of evaluating individual modalities, they are not always able to combine facial cues and textual material in a way that makes sense. The goal of this project is to provide a complete system that can efficiently combine and understand signals from textual data and facial emotions. To achieve a more robust and nuanced understanding of human emotions, the challenge involves not only developing accurate models for each modality but also developing an integrated framework that works in concert with both information sources to produce more sophisticated and reliable applications in areas like emotional analysis and human-computer interaction.

1.3 Objectives

The main aim of this project is:

- To create a Multimodal Emotion Recognition System from Text and Facial Expressions.

1.4 Scopes and Applications

The Multimodal Emotion Detection System can widely be used in various sectors for various purposes. Some of its application areas are mentioned below:

- **Human-Computer Interaction (HCI):** Multimodal emotion detection systems can enhance HCI by enabling devices and interfaces to adapt to users' emotional states. For example, smart assistants or chatbots equipped with emotion detection capabilities can respond more empathetically or tailor recommendations based on users' emotional cues.

- **Healthcare and Well-being:** Multimodal emotion detection systems can assist in mental health monitoring and therapy sessions. They can help therapists and caregivers track patients' emotional states over time, providing insights into mood fluctuations and potential indicators of conditions like depression or anxiety.

- **Education and Learning:** In educational settings, multimodal emotion detection systems can support personalized learning experiences. By analyzing students' facial expressions and textual responses, educators can gauge their levels of engagement, frustration, or comprehension.

- **Security and Surveillance**: This system can widely be used on airport security or crowd monitoring. By analyzing facial expressions and text inputs, these systems can help identify suspicious behavior or individuals displaying signs of distress or agitation.

- **Customer Service and Emotions Analysis:** By analyzing customers' facial expressions and textual feedback during interactions, businesses can better understand their needs and emotions, enabling more empathetic and effective responses.

Chapter 2

Literature Review

At present, the research on multimodal emotion recognition is a hot topic in the interdisciplinary research of cognitive science, physiology, linguistics, computer science, and so on. Multimodal emotion recognition has attracted more and more attention from scientific research institutions and researchers domestically and internationally. Emotion detection has gained significant attention in recent years due to its applications in various fields such as human-computer interaction, affective computing, and mental health. Traditional approaches have primarily focused on either text or facial expression analysis independently. However, the integration of multiple modalities, particularly text and facial expressions, has emerged as a promising avenue for improving the accuracy and robustness of emotion detection systems.

2.1 A Survey of Deep Learning-Based Multimodal Emotion

The paper by Hailun Lian et al. [1] provides a comprehensive survey of deep learning-based multimodal emotion recognition. It suggests using the Wav2Vec approach on the IEMOCAP and MELD datasets for feature extraction from vocal. In the textual domain, Bert and Roberta's models were used. For extracting features from facial expressions, 3D-CNN, OpenFace, and DenseNet were used. Concerning fusion strategies, Fine-grained Interaction Fusion was used to foster detailed interactions between modalities based on nuanced features.

2.2 Multimodal Emotion Recognition Model Based on a Deep Neural Network with Multiobjective Optimization

The paper by Mingyong Li et al. [2] proposes a multimodal emotion recognition model based on a deep neural network with multiobjective optimization. The proposed model takes features from audio and facial expressions. DeepCNN with decision level fusion based multiobjective algorithm was used. Traditional speech emotion recognition algorithms use LLDs or HSFs for feature extraction and HMM for emotion classification.

2.3 Speech Emotion Recognition Using Speech Feature and Word Embedding

The paper by Bagus Tris Atmaja et al. [3] focuses on speech emotion recognition using speech features and word embedding. The project focuses on IEMOCAP datasets and the ability to detect emotions from speech and textual words. In this paper, two unidirectional LSTM layers are used for text and fully connected layers are applied for acoustic emotion recognition. Both networks then are merged to produce one of four predicted emotion categories by fully connected networks.

2.4 Deep Emotion Recognition in Dynamic Data using Facial, Speech and Textual Cues: A Survey

The paper by Tao Zhang et al. [4] reviews different methods used in the field of multimodal emotion recognition (MER) in recent years. This paper comprehensively reviews and summarizes the definition of emotion models and the state-of-the-art of unimodal emotion recognition including facial expression recognition, speech emotion recognition, and textual emotion recognition in dynamic data.

2.5 Multimodal Emotion Recognition from Expressive Faces, Body Gestures and Speech

The paper by George Caridakis et al. [5] uses a multimodal approach for the recognition of eight emotions that integrates information from facial expressions, body movement, gestures, and speech. In this study, a Bayesian classifier uses a multimodal corpus with eight emotions and ten subjects. Data were fused at the feature level and the decision level.

Table 2.1: Multimodal Emotion Recognition Studies

SN	Title	Author/s	Methodology	Dataset	Accuracy
1	A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face [1]	Hailun Lian, Cheng Lu, Sunan Li, YanZhao, Chuangao Tang, Yuan Zong	3D-CNN, Wav2Vec, OpenFace, DenseNet, BERT, RoBERTa, Fine-grained interaction fusion	IEMOCAP and MELD	-
2	Multimodal Emotion Recognition Model Based on a Deep Neural Network with Multiobjective Optimization [2]	Mingyong Li, Xue Qiu, Shuang Peng, Lirong Tang, Qiqi Li, Wenhui Yang, Yan Ma	Deep CNN with decision level fusion based on a multiobjective algorithm	IEMOCAP	75.38%
3	Speech Emotion Recognition Using Speech Feature and Word Embedding [3]	Bagus Tris Atmaja, Kiyoaki Shirai, Masato Akagi	Word embedding, uni directional LSTM	IEMOCAP	75.49%
4	Deep Emotion Recognition in Dynamic Data using Facial, Speech and Textual Cues: A Survey [4]	Tao Zhang, Zhenhua Tan	-	-	-
5	Multimodal emotion recognition from expressive faces, body gestures and speech [5]	George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, Kostas Karpouzis	Feature level and Decision level fusion	GEMEP corpus	74.6%

Chapter 3

Requirement Analysis

3.1 Software Requirement

Software requirements for the prepared system include:

1. Google Colab
2. GitHub/GitLab
3. Kaggle
4. Keras
5. Jupyter Notebooks/Jupyter Lab
6. Matplotlib
7. NLTK (Natural Language Toolkit)
8. NumPy
9. OpenCV
10. Pandas
11. PyCharm
12. Python
13. PyTorch
14. RegEx
15. Scikit-learn
16. Seaborn
17. Streamlit
18. Tensorflow
19. Transformer
20. Vscode

3.2 Hardware Requirement

The project required the following hardware requirements:

1. NVIDIA TESLA 4 GPU
2. NVIDIA TESLA P100 GPU

3.3 Functional Requirement

3.3.1 Text Analysis

Ability to analyze text input for emotional content, including sentiment analysis and emotion classification. Support for various languages and dialects in text analysis.

3.3.2 Facial Expression Recognition

Detection and recognition of facial expressions from image frames.Identification of key facial features indicative of various emotions, such as happiness, sadness, anger, surprise, etc.

3.3.3 Integration of Text and Facial Analysis

Integration of text and facial analysis results to provide a more comprehensive understanding of emotional states.

3.4 Non-Functional Requirement

3.4.1 Accuracy and Reliability

High accuracy in detecting and classifying emotions from both text and facial inputs.Reliable performance across diverse demographic groups, cultural backgrounds, and language variations.

3.4.2 Performance

Efficient processing of text and facial inputs to provide timely responses.

3.4.3 Privacy and Security

Compliance with privacy regulations to ensure the protection of user data, especially facial images.

3.4.4 Adaptability

Flexibility to integrate with existing software systems or platforms, such as social media platforms, communication apps, or customer service tools.

3.5 Feasibility Study

The following points describe the feasibility of the project.

3.5.1 Economic Feasibility

Our project relies on open-source platforms and free resources, it offers scalability without incurring additional expenses. As the demand for emotion detection services grows, the system can easily scale up by provisioning additional resources without significant financial implications. Leveraging free GPUs or cloud-based services for computational tasks further minimizes infrastructure costs, as there is no expenditure on dedicated hardware or cloud computing resources. Maintenance and updates to the system can also be performed cost-effectively, as open-source communities often provide ongoing support and enhancements at no extra charge.

3.5.2 Technical Feasibility

The technical feasibility of multimodal emotion recognition, integrating facial expression and text-based approaches through machine learning, represents a potential future in the field of human-computer interaction. Leveraging modern algorithms and neural networks, this approach intends to synergize the strengths of visual and textual clues for a more thorough comprehension of human emotions. The facial expression component comprises the extraction of characteristics from face landmarks, applying computer vision algorithms to determine subtle subtleties in expressions. Simultaneously, the text-based approach focuses on natural language processing to evaluate written or spoken language, capturing the contextual complexities of emotions represented in words. Integrating these modalities boosts the model's robustness and adaptability across multiple communication channels. Challenges include data synchronization, model complexity, and real-time processing requirements. Despite these hurdles, the technological feasibility of multimodal emotion recognition offers enormous potential for applications in human-computer interaction, sentiment analysis, and mental health monitoring. Advances in machine learning approaches continue to fuel the progress of this interdisciplinary subject, paving the way for more nuanced and context-aware emotional intelligence in artificial systems.

3.5.3 Operational Feasibility

The operational feasibility of multimodal emotion detection using facial expression and text-based approaches with machine learning is promising. Multimodal emotion recognition is designed to use expression and speech information to recognize individual behaviors. Feature fusion can enrich various model information, which is an important way for multimodal emotion recognition. A learning-based method, M3ER, uses cues from multiple co-occurring modalities such as face, text, and speech, and is more robust than other methods to sensor noise in any of the individual modalities. It models a novel, data-driven multiplicative fusion method to combine the modalities, which learn to emphasize the more reliable cues and reduce others on a per-sample basis. Therefore, the operational feasibility of this method is high, given its robustness to sensor noise and its ability to successfully combine multiple modalities for improved emotion recognition.

Chapter 4

System Design and Architecture

A Multimodal Emotion Recognition System has been developed, which takes a facial expression and text as its input and processes them to provide emotional values as its output. The system can recognize seven emotions: anger, fear, sadness, happiness, disgust, neutrality, and surprise.

4.1 Use Case Diagram

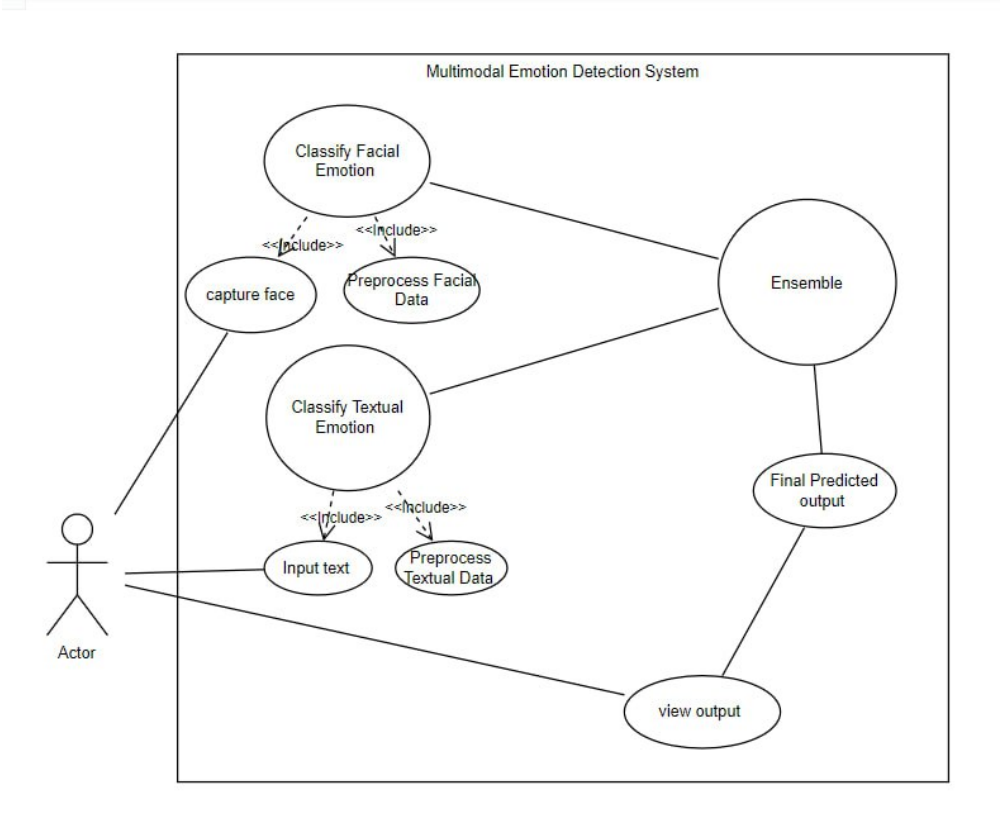

Figure 4.1: System Use Case Diagram

4.2 System Block Diagram

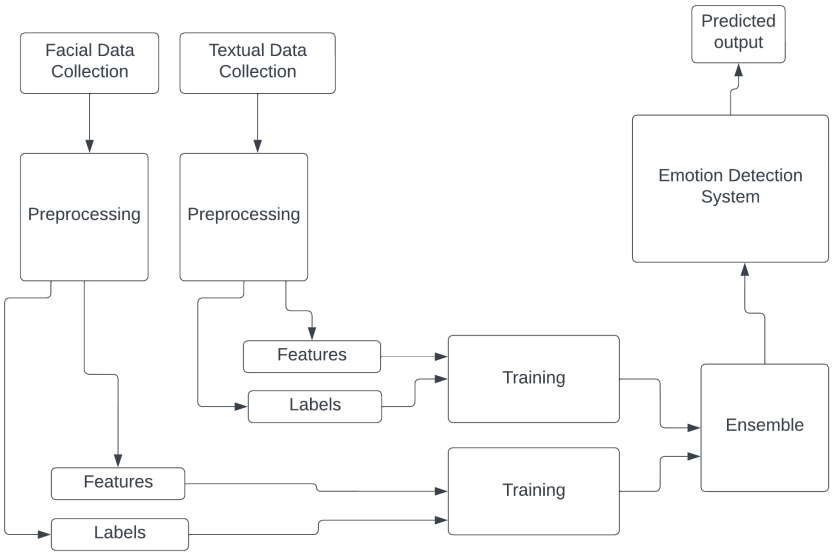

Figure 4.2: System Block Diagram for Training Pipeline

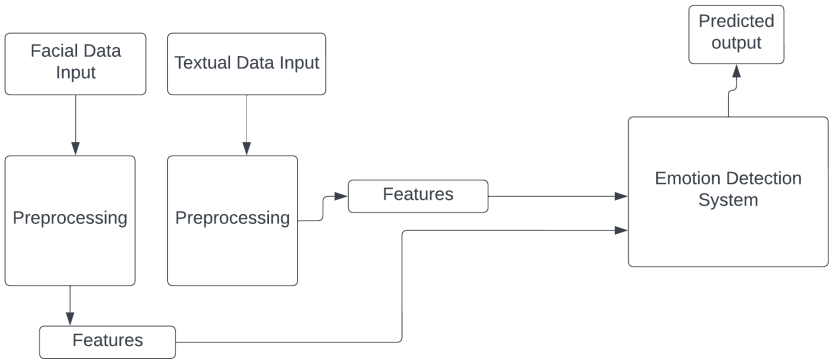

Figure 4.3: System Block Diagram for Inference Pipeline

4.3 Sequence Diagram

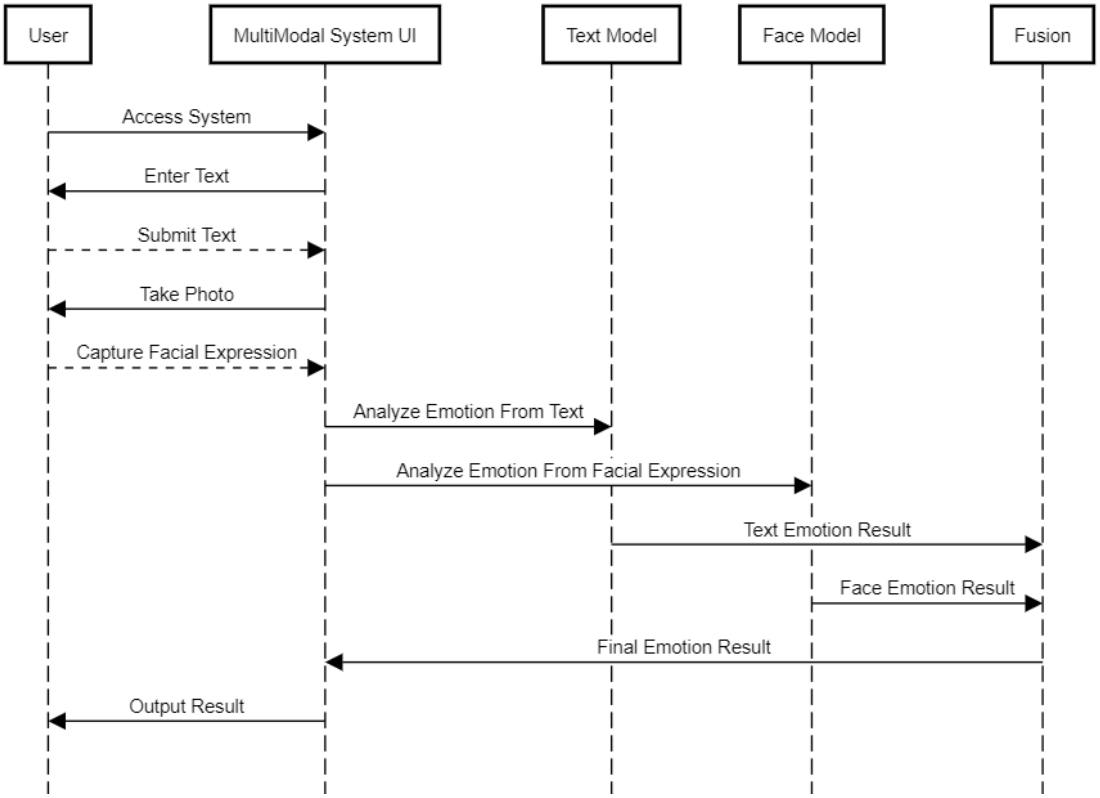

Figure 4.4: Sequence Diagram

Chapter 5

Methodology

5.1 Software Development Approach

The Prototype model is a software development approach wherein instead of freezing the requirements before design or coding can proceed, a throwaway prototype is built to understand the requirements. The prototypes are usually incomplete systems, with many details left out. The goal is to provide a system with overall functionality. In this model, the prototype of the actual system is created, the requirements are updated, and the system is rebuilt until the final requirements are met.

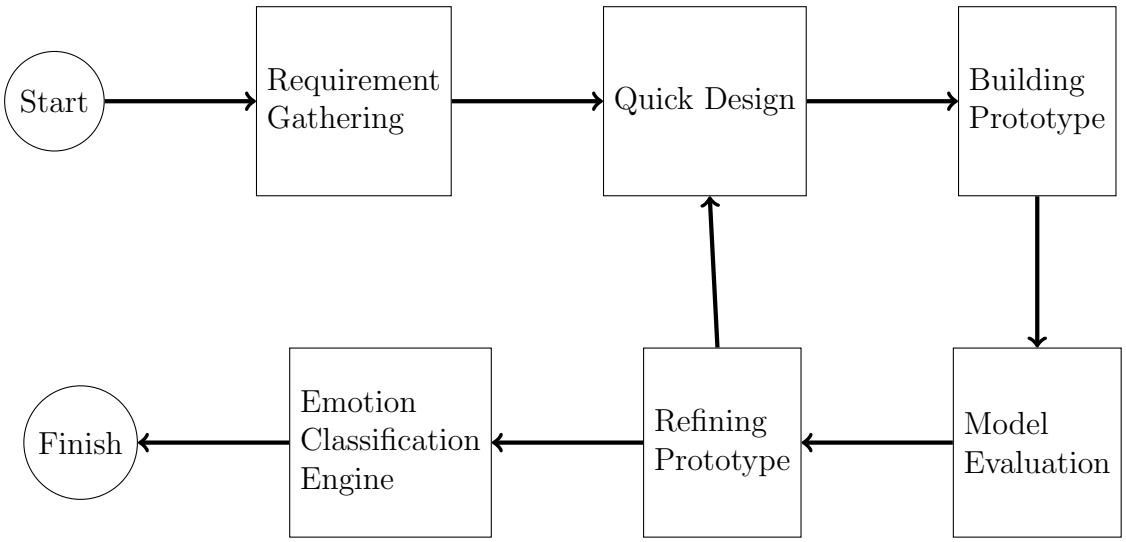

Figure 5.1: Prototype Model for Software Development

5.2 Data Collection

Hugging Face [6], Kaggle, and the MELD website provided us with textual and face datasets. Textual datasets have been collected from Hugging Face and the MELD dataset website, alongside facial datasets obtained from Kaggle. These datasets allow us for thorough analysis and modeling in both the visual and textual domains.
The datasets that were gathered and analyzed for our study are presented in the following table.

Table 5.1: Facial Datasets

S.N.	Dataset	Emotions	Samples	Description	Source
1	AffectNet dataset	anger, disgust, fear, happy, neutral, sad, surprise, and contempt	28,175 images	Each image is in grayscale and has a resolution of 96 x 96.	Kaggle
2	FER2013 (Facial Expression Recognition 2013)	anger, disgust, fear, happy, sad, surprise, and neutral	38,887 images	Each image is in grayscale and has a resolution of 48 x 48.	Kaggle

Table 5.2: Textual Datasets

S.N.	Dataset	Emotions	Samples	Description	Source
1	SetFit /emotion	sadness, anger, love, surprise, joy, and fear	20,000 text	It has properly labeled, cleaned lowercase sentences.	Hugging Face
2	Multimodal Emotion Lines Dataset (MELD)	anger, disgust, sadness, joy, neutral, surprise, and fear	1,400 dialogues and 13,000 utterances	Includes audio clips and dialogues from Family TV series.	MELD

5.2.1 AffectNet Dataset

The sample images from the AffectNet dataset are:

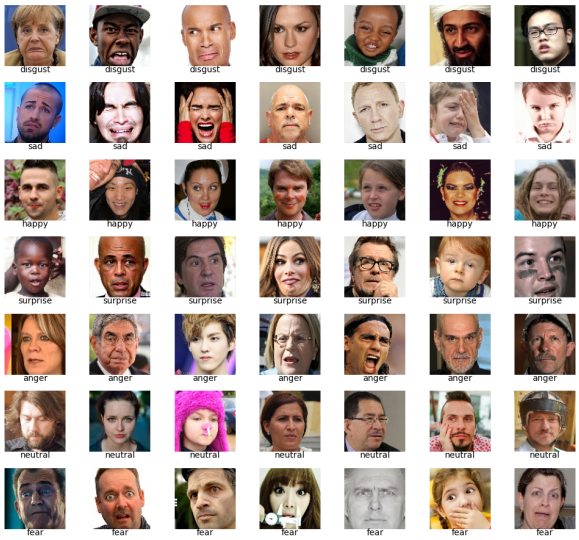

Figure 5.2: Image samples from AffectNet dataset

5.2.2 FER2013 Dataset

The sample images from the FER2013 dataset are:

Figure 5.3: Image samples from FER2013 dataset

5.2.3 SetFit/Emotion Dataset

The sample data of the SetFit/Emotion dataset are:

Table 5.3: SetFit/Emotion Dataset

Text	Label	Label Text
I think it's the easiest time of year to feel dissatisfied	3	Anger
I feel as confused about life as a teenager or as jaded as a year-old man	4	Fear
I do not feel reassured anxiety is on each side	1	Joy
I feel romantic too	2	Love
I feel like I have to make the suffering I'm seeing mean something	0	Sadness
I've been taking 40 milligrams or 8 times recommended amount and I've fallen...	5	Surprise

5.2.4 MELD Dataset

The sample data from the MELD dataset:

Table 5.4: MELD Dataset

Sr No.	Utterance	Speaker	Emotion	Sentiment
1	Got me.	Monica	Sadness	Negative
2	Can I get a beer.	Chandler	Neutral	Neutral
3	You betcha!	Chandler	Joy	Positive
4	I'm not even... I'm not even	Chandler	Neutral	Neutral

5.3 Data Preparation

5.3.1 For Facial Datasets

After acquiring facial datasets from FER2013 and AffectNet, several preprocessing steps were conducted to ensure they were in the appropriate format for neural network training. FER2013 comprises grayscale images with dimensions of 48x48 pixels, whereas AffectNet comprises color images sized at 96x96 pixels. As AffectNet includes additional images associated with 'contempt', a category absent in FER2013, all images linked to the 'contempt' emotion were omitted from the AffectNet dataset. Subsequently, the folders were structured, and the datasets were categorized appropriately to maintain uniformity. The finalized AffectNet

dataset comprises 26,171 images.

The statistics and distribution of emotion classes within the AffectNet dataset are as follows:

Table 5.5: AffectNet Dataset Statistics

Emotions	**Samples Count**
anger	3218
disgust	2477
fear	3176
happy	5044
neutral	5126
sad	3091
surprise	4039

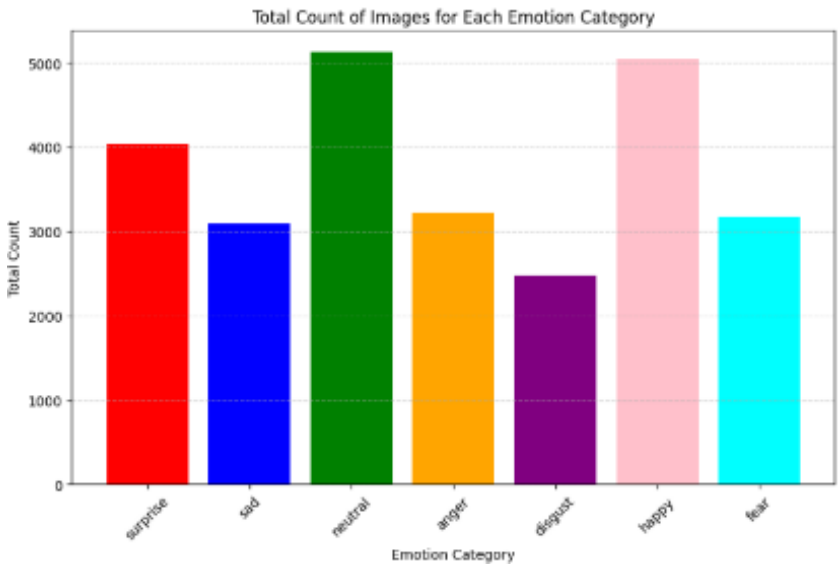

Figure 5.4: Data distribution across various emotions in AffectNet dataset

The statistics and distribution of emotion classes within the FER2013 dataset are as follows:

Table 5.6: FER2013 Dataset Statistics

Emotions	**Samples Count**
anger	4953
disgust	3547
fear	5121
happy	8989
neutral	6198
sad	6077
surprise	4002

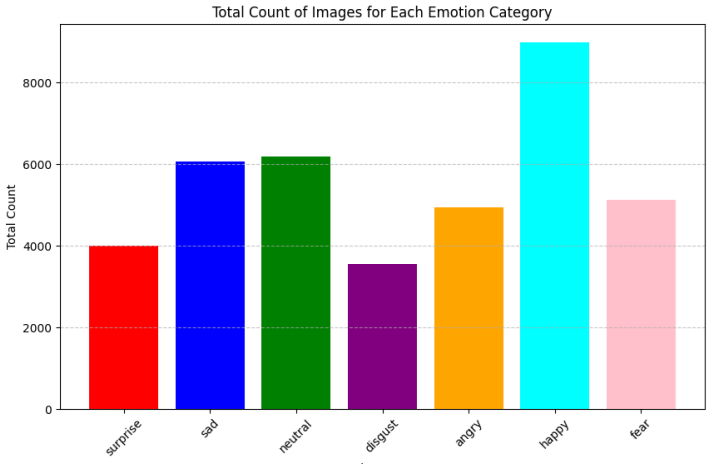

Figure 5.5: Data distribution across various emotions in FER2013 dataset

To prepare the facial data for neural network training and facilitate proper interpretation and learning, the following data augmentation and preprocessing steps were implemented:

1. **Data Augmentation:** Following data-augmentation techniques are applied to increase the diversity of the model and to enhance the model's ability to generalize to unseen data.

(a) **Horizontal flip**

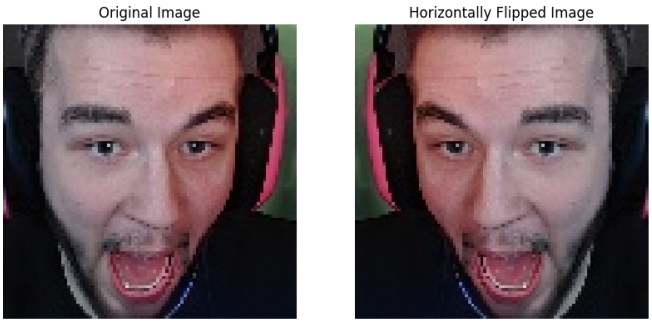

Figure 5.6: Image Horizontal Flip

(b) **Rotation:** rotating the image within a range of 20°.

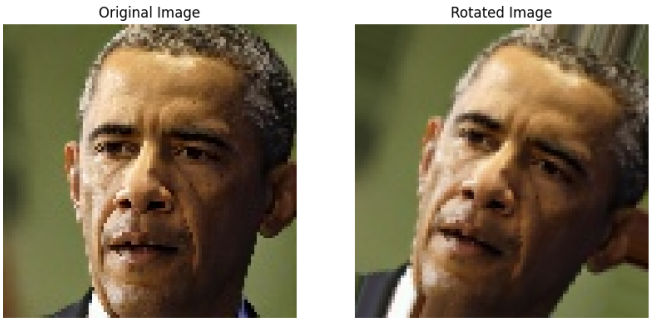

Figure 5.7: Image Rotation

(c) **Width shift:** shifting the image within a range of 0.1 by width.

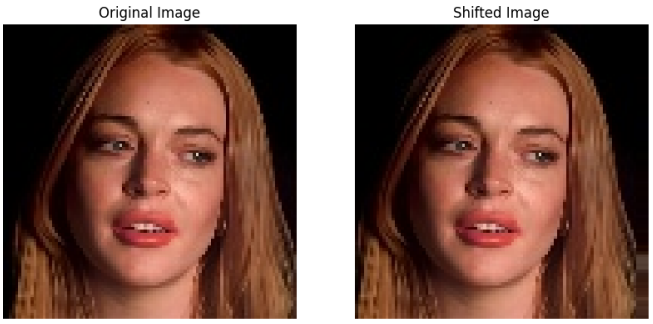

Figure 5.8: Width Shift

(d) **Height shift:** shifting the image within a range of 0.1 by height.

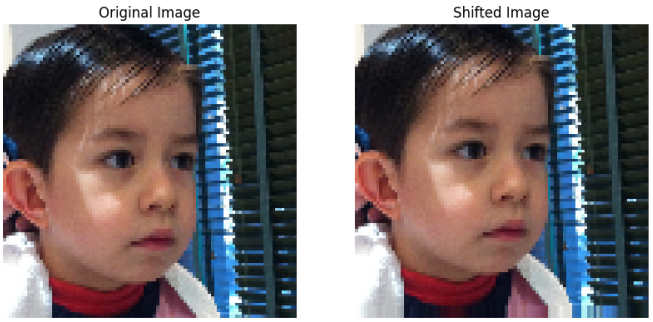

Figure 5.9: Height Shift

2. **Grayscale conversion:** No additional color processing was required for the grayscale photos from FER2013. To maintain consistency with the FER2013 format, the color photos from AffectNet were converted to grayscale.

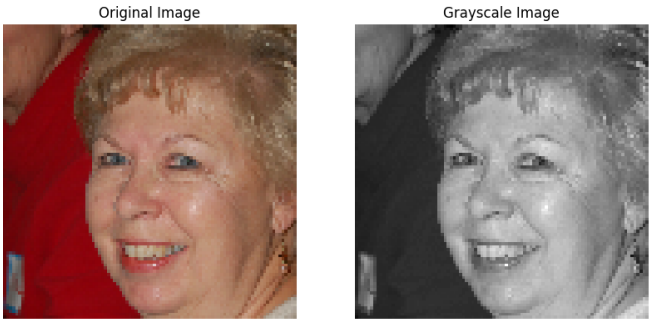

Figure 5.10: Grayscale Conversion

3. **Resizing:** The images in the FER2013 dataset were resized from 48x48 pixels to 56x56 pixels. This process involved adjusting the dimensions of each image and increasing its size.

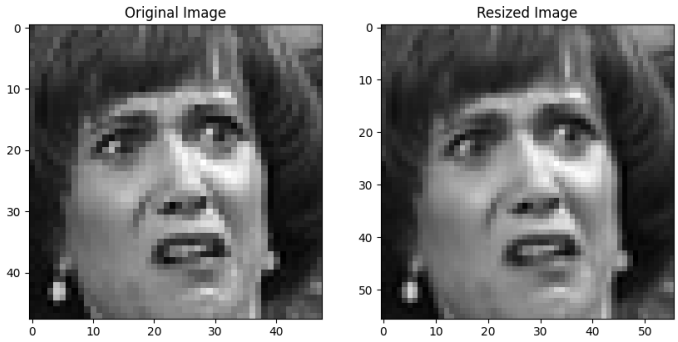

Figure 5.11: Resizing Image

4. **Normalization:** To improve model performance and enable more effective convergence during training, the pixel values of all the images were normalized to a range of [0, 1].

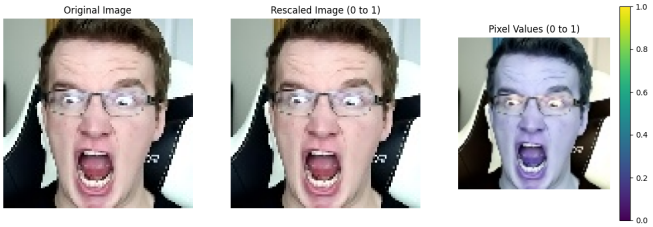

Figure 5.12: Normalizing Image

5.3.2 For Textual Datasets

The emotions available in the SetFit/emotion dataset were found to be incomplete for the analysis, as it includes additional text related to the 'love' emotion but lacks text related to 'neutral' and 'disgust' emotions. Furthermore, upon observation, it was noted that the label 'joy' in the dataset closely corresponds to 'happy', thus it was renamed accordingly. To address these deficiencies, text related to 'neutral' and 'disgust' from the MELD dataset was integrated, and additional text was generated with the assistance of ChatGPT. Subsequently, this supplementary data was manually combined with the existing SetFit/emotion dataset to construct a textual dataset tailored specifically for emotion recognition.
In the data preprocessing pipeline, uncleaned text was identified in the MELD dataset, containing HTML tags, URLs, and symbols. To address these challenges, preprocessing steps were performed with the help of the regular expression module.

1. **Removal of html tags:**
   I am `<b>extremely</b> <i>angry</i>` right now → I am extremely angry right now.
2. **Removal of urls:**
   I love this website: `https://www.facebook.com` → I love this website.
3. **Removal of other symbols and emojis:**
   ↓temperature is → 36°C ☺→ temperature is 36
4. **Removal of punctuation:**
   No, don't , I don't want to see you. → No dont I dont want to see you.
5. **Lower casing the sentences:**
   Accidentally left the Caps lock on → accidentally left the caps lock on

After completing these preprocessing procedures, a final text dataset that had been cleaned was obtained. This dataset combines the carefully processed information from the MELD dataset with the refined data from the SetFit/emotion dataset.

Table 5.7: Final Textual Dataset

Text	Label	Label Text
i am feeling grouchy	2	Anger
their gossiping behind others backs bothers me...	5	Disgust
i feel so shaken and guilty for not being a be...	3	Fear
i feel more lively	1	Happy
unfortunately for me you have to be 23 or olde...	6	Neutral
im feeling kind of unwelcome	0	Sad
i admire makes me feel amazed at my life	4	Surprise

The statistics and distribution of emotion classes within the final dataset are as follows:

Table 5.8: Final Textual Dataset Statistics

Emotions	Samples Count
anger	3886
disgust	1262
fear	2373
happy	6761
neutral	4907
sad	5797
surprise	719

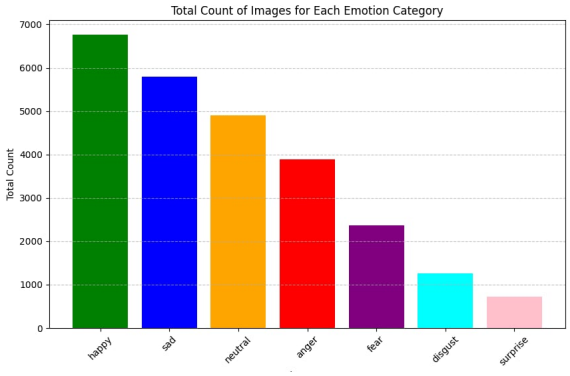

Figure 5.13: Data Distribution Across Various Emotions in Final Textual Dataset

5.4 Training

The training process is started after the division of each dataset into train, validation, and test sets with a split ratio of 0.8, 0.1, and 0.1 respectively. The data will be fed into the training system or neural network as shown in the figure below:

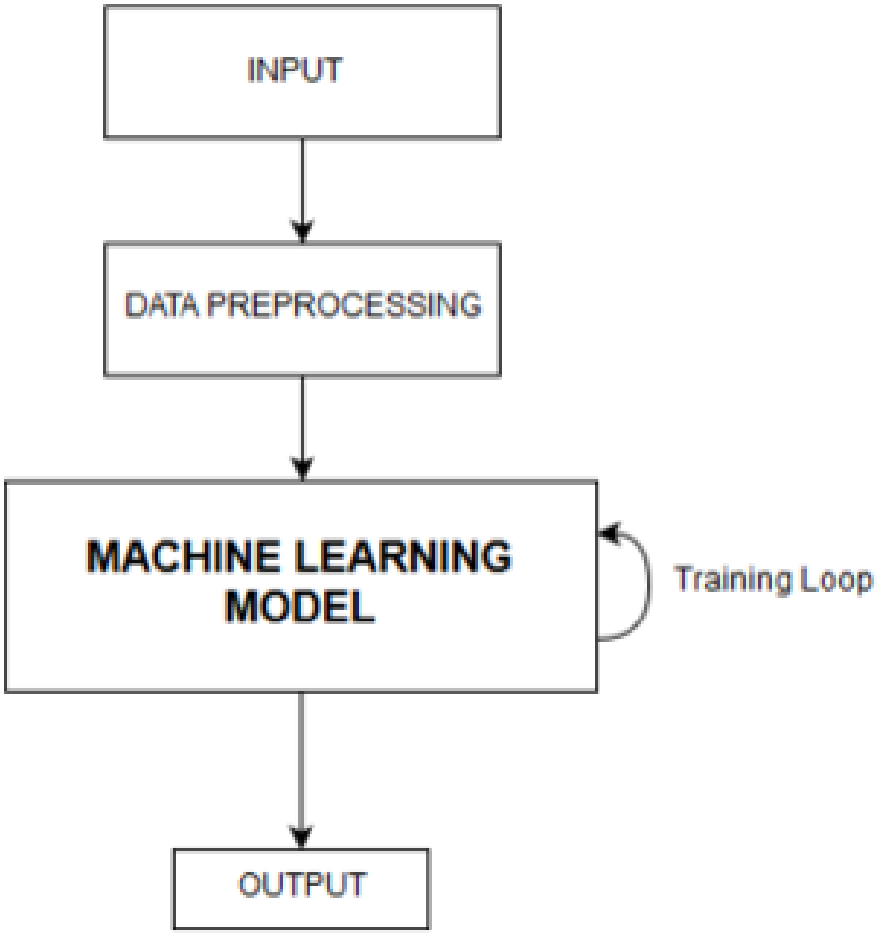

Figure 5.14: Basic Machine Learning Training Process

5.5 Models and Algorithms

The various algorithms and models were developed for building the emotion recognition model. The following topics describe all the algorithms and models used in the project in detail.

5.5.1 CNN Model

Convolutional neural network models [7], or CNN models, are specialized neural network architectures designed to analyze structured input in the form of grids, especially images. A CNN model learns to extract hierarchical features from input images by arranging convolutional, activation, pooling, and fully connected layers hierarchically. This allows the model to perform specialized tasks like object identification, image segmentation, and image classification. Through iteratively adjusting its parameters using backpropagation and optimization methods to minimize the gap between its predictions and the ground truth, the model learns to map input images to corresponding labels or annotations during training. The CNN model is a flexible and effective tool in computer vision applications because, once trained, it can reliably identify items inside images, locate objects and their positions, or split images into meaningful sections.

5.5.1.1 CNN Architecture

One common architecture in Deep Learning, especially in Computer Vision, is the convolutional neural network (CNN). Computer vision is the branch of artificial intelligence that deals with giving computers the ability to see, understand, and interpret images or visual data. Each of the several layers that make up a CNN—the input, convolutional, pooling, and fully connected layers, for example—is essential to the processing of visual data.

Within the CNN architecture, the Pooling layer minimizes the picture dimensions to aid in computational efficiency, while the Convolutional layer applies filters to the input image to efficiently extract key features. Ultimately, the completely connected layer is in charge of producing the most accurate forecasts. The network improves its capacity to identify patterns in visual data by optimizing its filters through techniques like gradient descent and backpropagation.

5.5.1.2 Custom ConvNet Model

For classification, a CNN architecture is utilized, comprising four convolutional layers, two fully connected (dense) layers, and an output layer, along with other essential components. The convolutional layers' main goal is to gradually extract features from the input images. The goal of the two fully connected (dense) layers is to further process the features extracted by the convolutional layers and make them suitable for final classification. Softmax activation is used by the seven neurons that make up the output layer to calculate class probabilities. For efficient picture classification, the model integrates convolutional and dense layers, batch normalization, activation functions, dropout, and max pooling. It is compiled

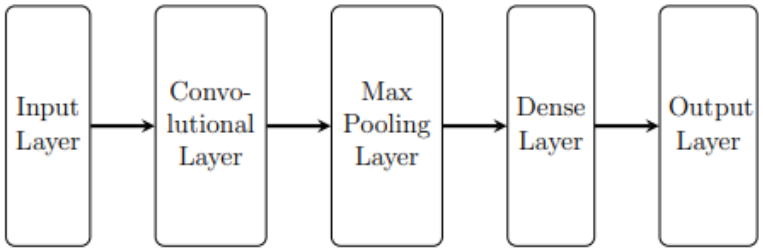

Figure 5.15: Simple CNN Architecture

with the Adam optimizer (lr=0.0001), categorical cross-entropy loss, and accuracy metric. The model along with hyperparameters is shown below:

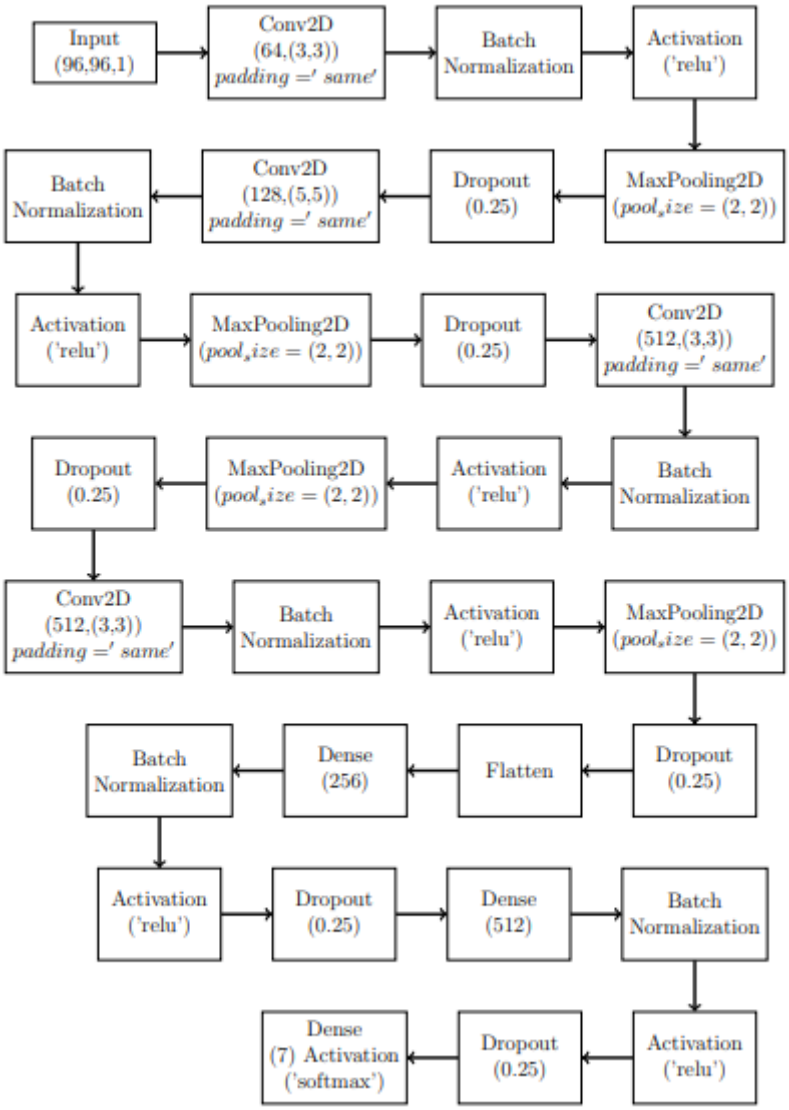

Figure 5.16: Custom Classification Model - ConvNet

This model has achieved a training accuracy of 69.63%, validation accuracy of 70.29%, training loss of 0.79 validation loss of 0.77 after training 80 epochs in Google Colab GPU kernel.

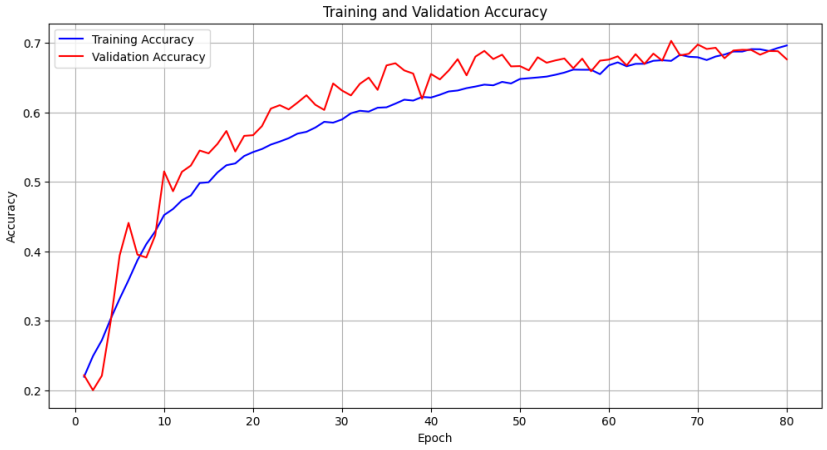

Figure 5.17: Learning Curve: Accuracy Plot for Custom ConvNet

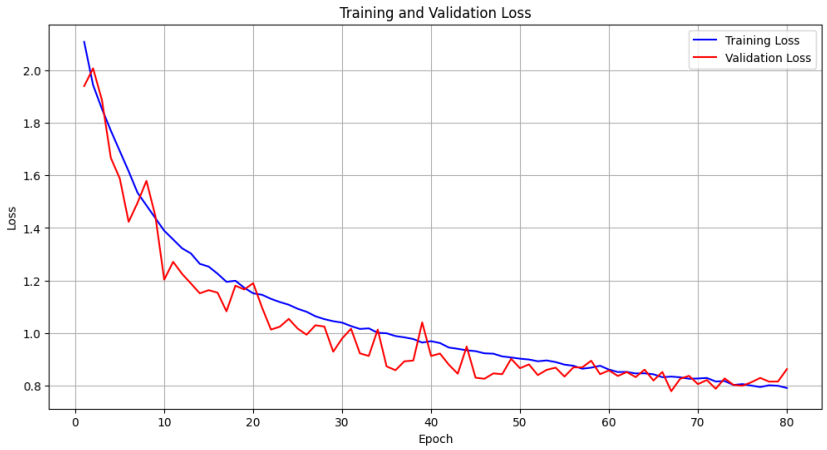

Figure 5.18: Learning Curve: Loss Plot for Custom ConvNet

5.5.2 BERT Model

BERT (Bidirectional Encoder Representations from Transformers) [8] is a state-of-the-art pre-trained language representation model developed by Google. It employs a bidirectional approach, processing input text simultaneously in both directions to capture context from preceding and following words. Built upon the transformer architecture, BERT uses self-attention mechanisms to focus on different parts of the input text, enabling it to understand complex relationships between words. Pre-trained on large-scale text corpora using tasks like Masked Language Modeling and next-sentence prediction, BERT learns deep contextualized representations of words. These pre-trained representations can be fine-tuned on specific downstream tasks with relatively small datasets, making BERT highly versatile and effective for a wide range of natural language processing tasks. BERT utilizes a stack of transformer encoder layers for this purpose. Here's an overview of the encoders in BERT models:

1. **Transformer Architecture**

   - BERT models are based on the transformer architecture, which consists of multiple layers of self-attention mechanisms and feedforward neural

networks.

- Each transformer encoder layer independently processes the input tokens in parallel, allowing BERT to capture complex dependencies and relationships within the input sequence.

2. **Self-Attention Mechanism**

   - The self-attention mechanism in each encoder layer enables BERT to weigh the importance of each token in the input sequence based on its context.
   - Tokens attend to each other, allowing the model to capture long-range dependencies and learn contextual representations of tokens.

3. **Feedforward Neural Network**

   - After the self-attention mechanism, each token representation passes through a feedforward neural network within the encoder layer.
   - The feedforward network applies non-linear transformations to the token representations, allowing the model to capture complex patterns and relationships within the data.

4. **Layer Normalization**

   - Layer normalization is applied after each sub-layer (self-attention and feedforward network) within the encoder layer.
   - Layer normalization normalizes the activations of the tokens within each layer, improving the stability and convergence of the model during training.

5. **Residual Connections**

   - BERT employs residual connections around each sub-layer within the encoder layer.
   - Residual connections allow the model to learn residual representations, making it easier to propagate gradients through the network and mitigate the vanishing gradient problem.

6. **Stacking of Encoder Layers**

   - BERT models typically consist of multiple stacked encoder layers.
   - The stacking of encoder layers allows BERT to capture hierarchical and multi-level representations of the input text, enabling it to learn rich contextualized embeddings for downstream tasks.

Overall, the encoders in BERT models play a crucial role in processing the input tokens, capturing contextual information, and generating representations that are rich in semantic meaning and context. These representations are then used as input to downstream tasks such as classification, regression, or sequence labeling.

5.5.3 Fined Tuned BERT Model

For classification, the pre-trained model 'bert-base-uncased' and its corresponding tokenizer were imported from the Hugging Face Transformers library. Subsequently, the text was tokenized with padding and truncation, and the train, validation, and test datasets were encoded. Next, the labels were associated with the input_ids, attention mask, and token type IDs. These data were then fed to the BERT model along with an output layer, whose activation function is 'softmax', producing probabilities as an output. With a learning rate of 1e-5 and sparse categorical cross-entropy loss, the model was compiled using the Adam optimizer. To address data imbalance challenges, a weighted loss function approach was utilized. The classification model using BERT is outlined below:

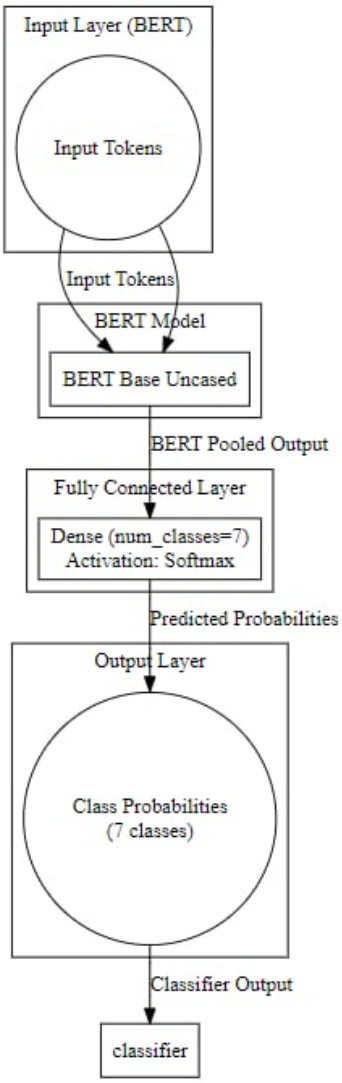

Figure 5.19: Fined Tuned BERT Model

This model has achieved a training accuracy of 92.95% , validation accuracy of 92.46%, training loss of 0.175 validation loss of 0.181 after training 3 epochs in the Google Colab GPU kernel.

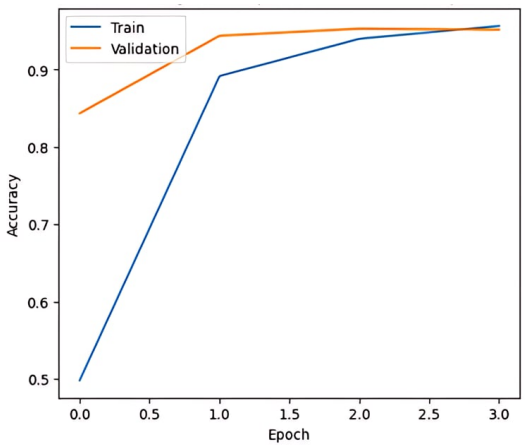

Figure 5.20: Learning Curve: Accuracy Plot for Fined Tuned BERT Model

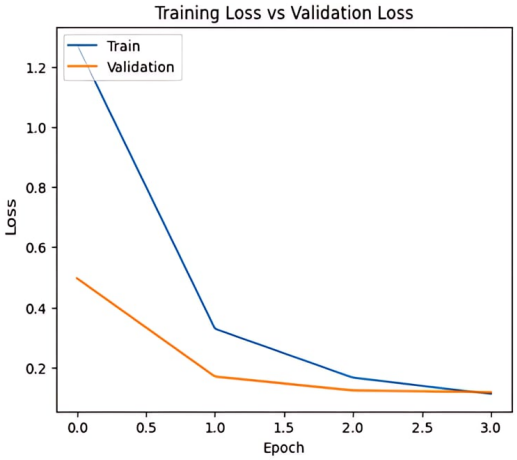

Figure 5.21: Learning Curve: Loss Plot for Fined Tuned BERT Model

5.5.4 Tokenization Algorithm

The BERT tokenizer's tokenization algorithm consists of a sequence of steps that transform unprocessed text input into numerical representations that can be entered into a BERT model. First, a method known as WordPiece tokenization divides the input text into individual words or subwords. This entails dissecting words into more manageable, meaningful chunks, particularly for words that are not in common usage. The next step involves mapping each token to an index in the model's vocabulary so that it can be used as an input ID. To indicate the start and finish of the input sequence, additional tokens are added, such as [CLS] (classification) and [SEP] (sentence separator). Padding tokens may also be added to guarantee a consistent sequence length. In the end, an attention mask is produced to show which tokens are padding tokens and which actually match words in the input. Through the tokenization process, the input text is made correctly formatted and transformed into a format that the BERT model can process for a variety of natural language processing tasks. Thus, to preprocess text data for BERT models, the presented tokenization algorithm combines both word-level and sentence-level tokenization techniques.The algorithm is given below:

BERT Tokenization Algorithm

1. Start
2. Input Sentences
3. Tokenization (WordPiece)
4. Convert Tokens to Input IDs
5. Add Special Tokens ([CLS], [SEP])
6. Add Padding Tokens (if necessary)
7. Generate Attention Mask
8. End

Table 5.9: BERT Input Details

Input Sentence	*Happiness filled her heart.*
Field	**Value**
Tokens	['[CLS]', 'happiness', 'filled', 'her', 'heart', '.', '[SEP]', '[PAD]', '[PAD]', '[PAD]']
Input IDs	[101, 8404, 3561, 2014, 2540, 1012, 102, 0, 0, 0]
Special Tokens	['[CLS]', '[SEP]', '[PAD]']
Padding Token	['[PAD]']
Attention Mask	[1, 1, 1, 1, 1, 1, 1, 0, 0, 0]

5.5.5 Ensemble Algorithm

The algorithm of ensemble [9] learning is:

1. **Predictions from Facial Model:**
   - Use the facial model to predict emotions for a given face.
   - Get the probabilities of each emotion class from the facial model.
2. **Predictions from Textual Model:**
   - Use the textual model to predict emotions for a given sentence.
   - Get the probabilities of each emotion class from the textual model.
3. **Combine Predictions:**
   - Ensure both models produce predictions for the same set of classes.
   - For each emotion class:
     - Take the average of the probabilities predicted by the facial and textual models.
     - Store the combined probability for each emotion class.
4. **Finding Predicted Emotion:**
   - Find the emotion with the highest combined probability among all classes.
   - This emotion is selected as the predicted emotion.

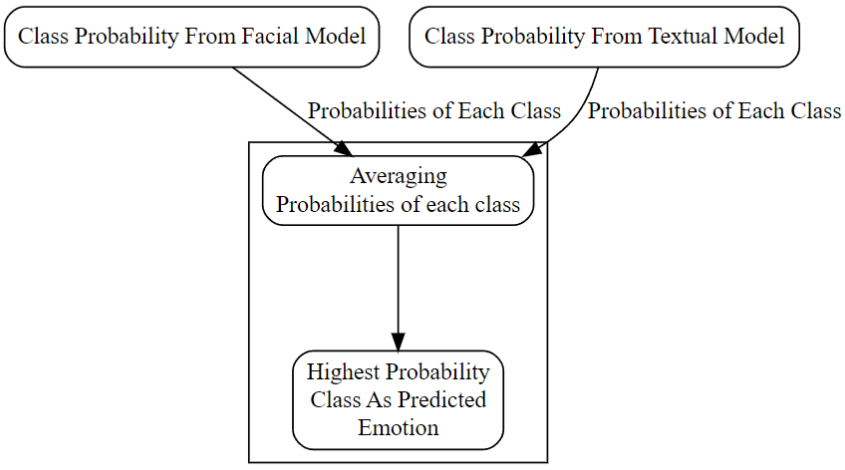

Figure 5.22: Ensemble Algorithm

Chapter 6

Result and Analysis

The system for emotion detection from facial and textual data has been completed. The input data from the file was successfully retrieved, then preprocessed and sent to the model for training.
For the Custom ConvNet model to detect the emotions from the facial data various datasets were used to train our custom model for facial emotion detection and the results of training those datasets on the model are shown in the table below:

Table 6.1: Dataset Metrics

Dataset	Training Accuracy	Validation Accuracy	Epoch
FER2013	64.58%	64.06%	80
AffectNet	69.63%	70.29%	80

The AffectNet Dataset was found to be best for the model.

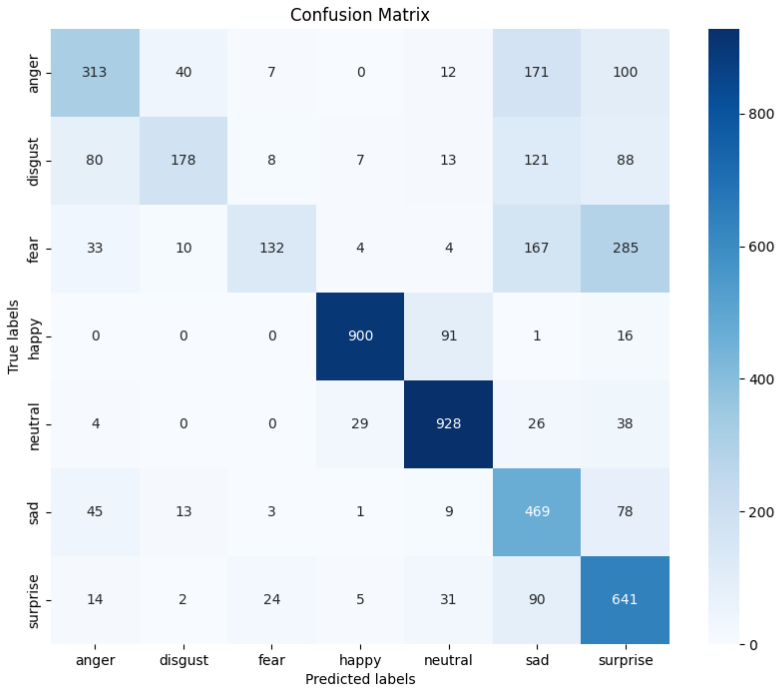

Figure 6.1: Confusion Matrix of the Custom ConvNet Model

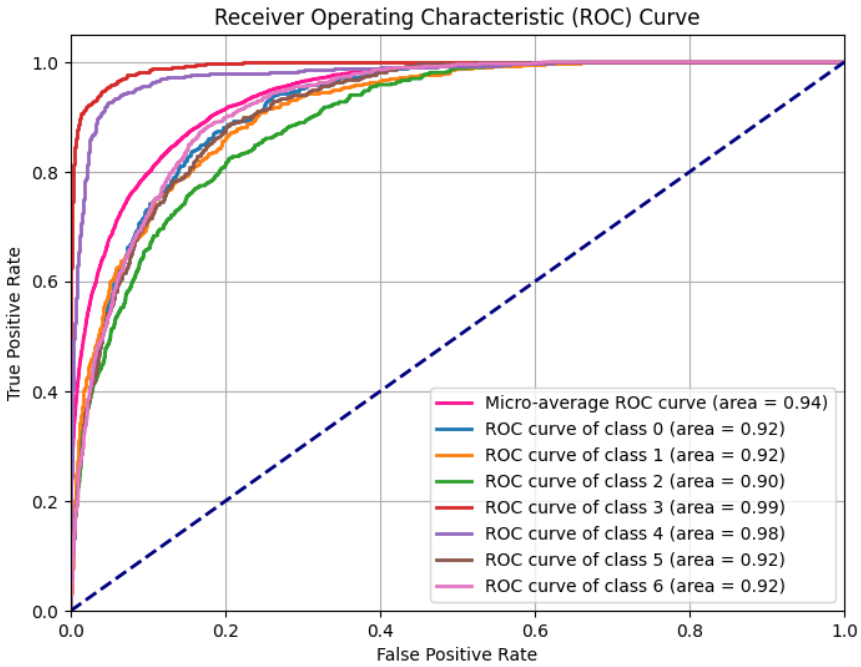

Figure 6.2: ROC Curve of the Custom ConvNet Model

Table 6.2: Performance Metrics of the Custom ConvNet Model

Parameters	**Values**
Accuracy	0.6807
Precision	0.7229
Recall	0.6807
F1 Score	0.6658
ROC AUC Score (Micro-average)	0.9348

After training the model on the AffectNet dataset, attempts were made to increase its accuracy by adjusting parameters such as the learning rate and filter size. To address overfitting issues, various measures were employed, including dropout and batch normalization. Additionally, early stopping techniques were incorporated into the training process.

For emotion detection from text, the pre-trained BERT model was employed. BERT, being a large model for text classification, has already been trained on a vast amount of data. To adapt it to our specific task, the datasets were fine-tuned by passing them through the BERT model. The final dataset obtained from Hugging Face was used for fine-tuning the model.
The results of the model are shown in the following figures :

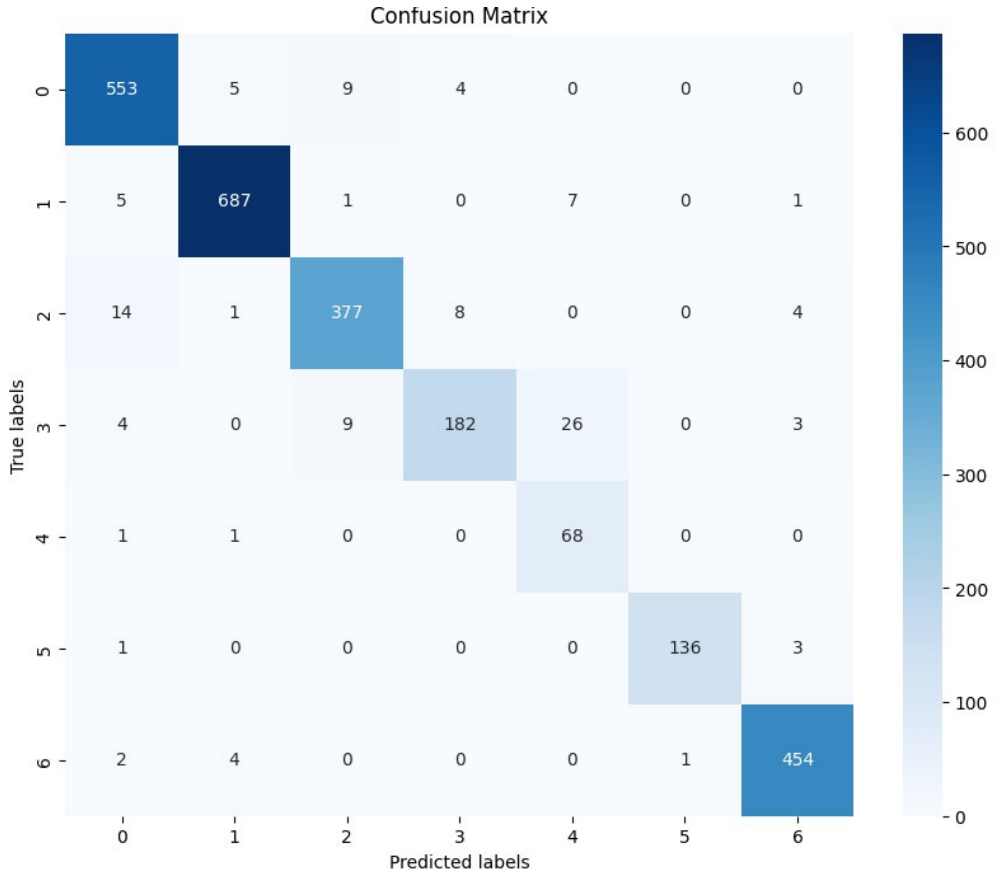

Figure 6.3: Label Information

Table 6.3: Final Textual Dataset Statistics

Label Text	**Label**
anger	2
disgust	5
fear	3
happy	1
neutral	6
sad	0
surprise	4

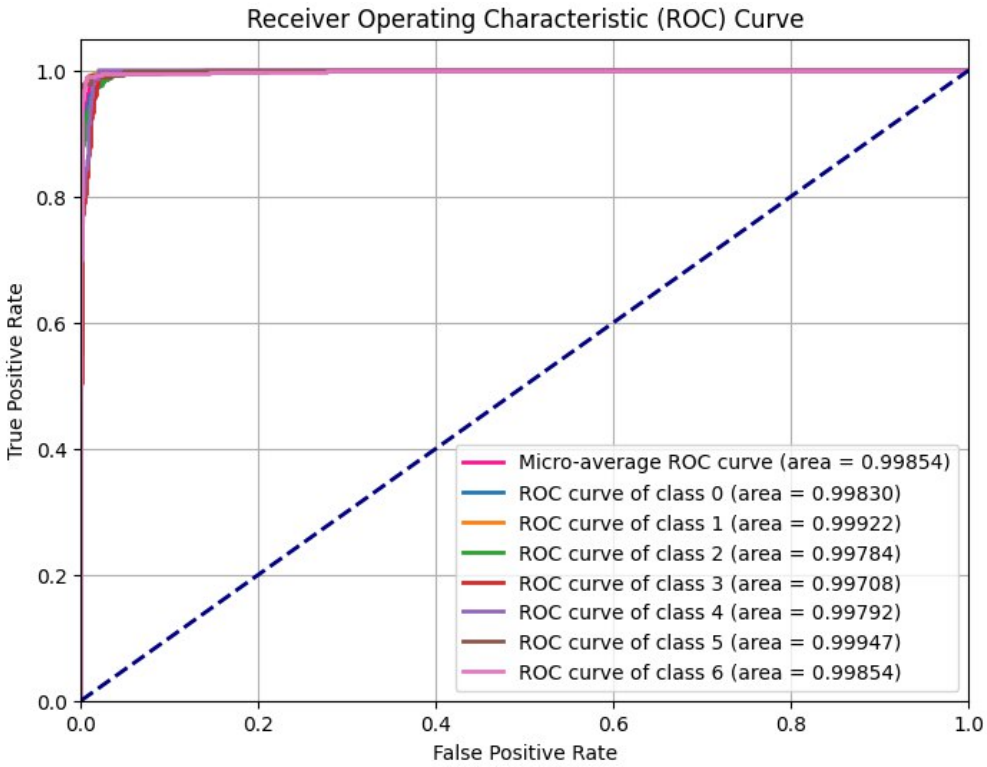

Figure 6.4: ROC Curve of Fined Tuned BERT model

Table 6.4: Performance Metrics of the Fined Tuned BERT Model

Parameters	**Values**
Accuracy	95.5659%
Precision	0.9589
Recall	0.9556
F1 Score	0.9561
ROC AUC Score	0.9983

The above results were obtained from fine tuning the BERT model with our datasets. Similarly, the inference system also works well. It successfully sends the user input to the machine learning model and then responds with the emotions that the input is. The screenshot of the output is shown in Outcomes section.

Chapter 7

Outcome

A simple user interface was created to interact with our model and predict the emotion through it. It has a space to enter the text from the user and take a photo to analyze the emotion in the text and the photo. Both text and photo are passed to the model and the result is again shown in the interface with the meter bars of each emotion included.

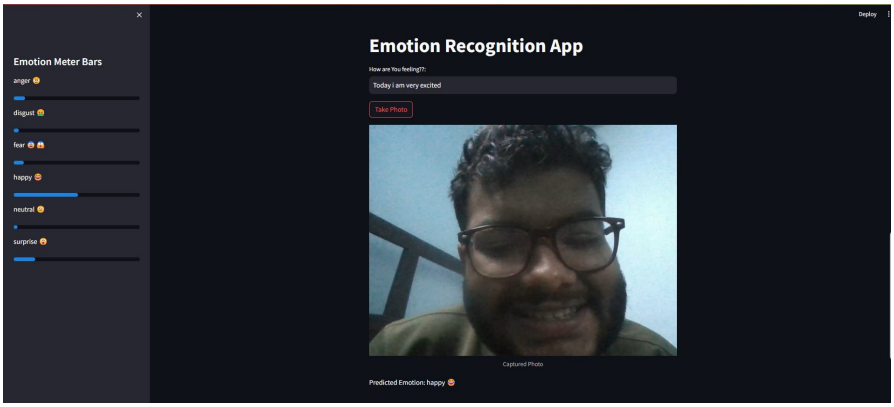

Figure 7.1: Interface of the Multimodal System

Chapter 8

Conclusion and Future Enhancements

8.1 Conclusion

The research aims to integrate facial expression and textual content processing for emotion recognition. By employing machine learning techniques on multimodal data, which includes facial features and text, improved accuracy and robustness in emotion recognition models are anticipated. The study's outcomes have the potential to advance applications in human-computer interaction, affective computing, and related fields, ultimately leading to the development of more intuitive and emotionally intelligent systems.

8.2 Limitations

Limitations of our project are:

- The accuracy of the Custom ConvNet model is lower than expected.
- When two distinct emotions are sent as input, the multimodal is unable to prioritize between two emotions.
- The accuracy of the system as a whole could not be determined as ensemble averaging was employed from the obtained output of both the facial and textual models. However, the accuracy of the two distinct models alone was examined.

8.3 Future Enhancements

Future enhancements to be made in our system are:

- Gather more data to increase the accuracy of the Custom ConvNet model.
- Implement feature concatenation and fusion model concepts.
- Collaboratively pre-train the multimodal system using a dataset containing textual and facial data.
- Embed audio mode into the system.

Bibliography

[1] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: speech, text, and face," *Entropy*, vol. 25, no. 10, p. 1440, Oct. 2023.

[2] C. I. A. Neuroscience, "Retracted: Multimodal sensor motion intention recognition based on three-dimensional convolutional neural network algorithm," *Computational Intelligence and Neuroscience*, vol. 2023, p. 1, Jun. 2023.

[3] "Speech emotion recognition using speech feature and word embedding," in *IEEE Conference Publication — IEEE Xplore*, Nov. 01 2019.

[4] T. Zhang and Z. Tan, "Deep emotion recognition using facial, speech and textual cues: A survey," INDIGO (University of Illinois at Chicago), Oct. 2021.

[5] G. Caridakis *et al.*, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *Springer eBooks*, 2007, pp. 375–388.

[6] Hugging Face. (n.d.) Hugging Face. Accessed: INSERT DATE. [Online]. Available: https://huggingface.co/

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, iSBN: 978-0262035613.

[8] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," *arXiv preprint arXiv:1810.04805*, 2018.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning for generic object recognition: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 109–139, 2013.

Appendix

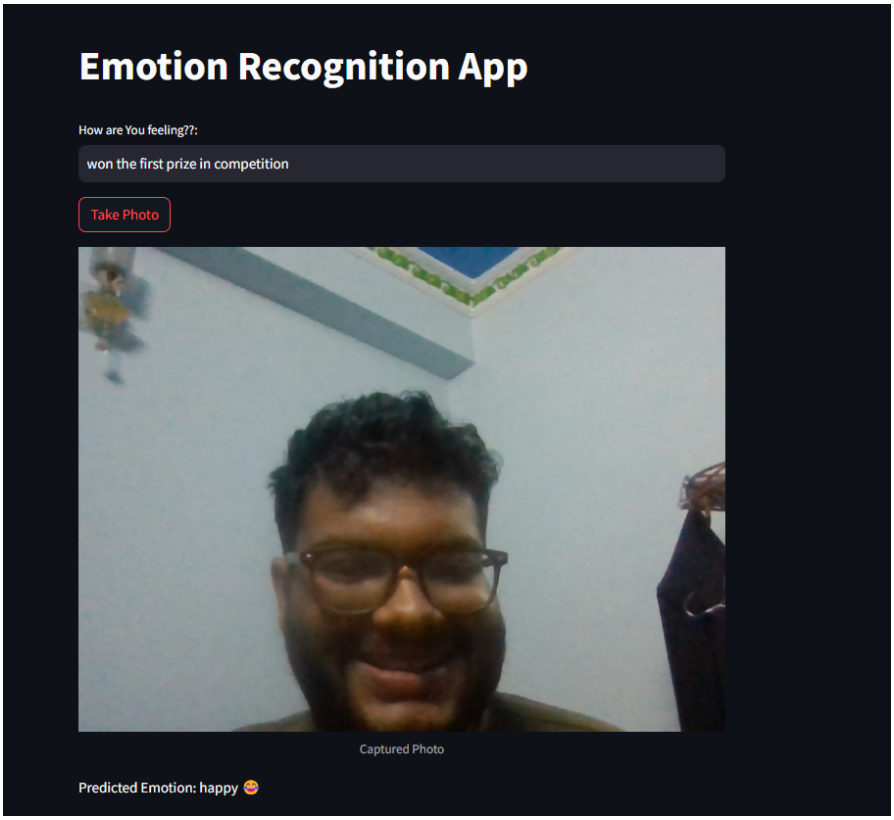

Figure A-1: True Prediction by the System: Output 1

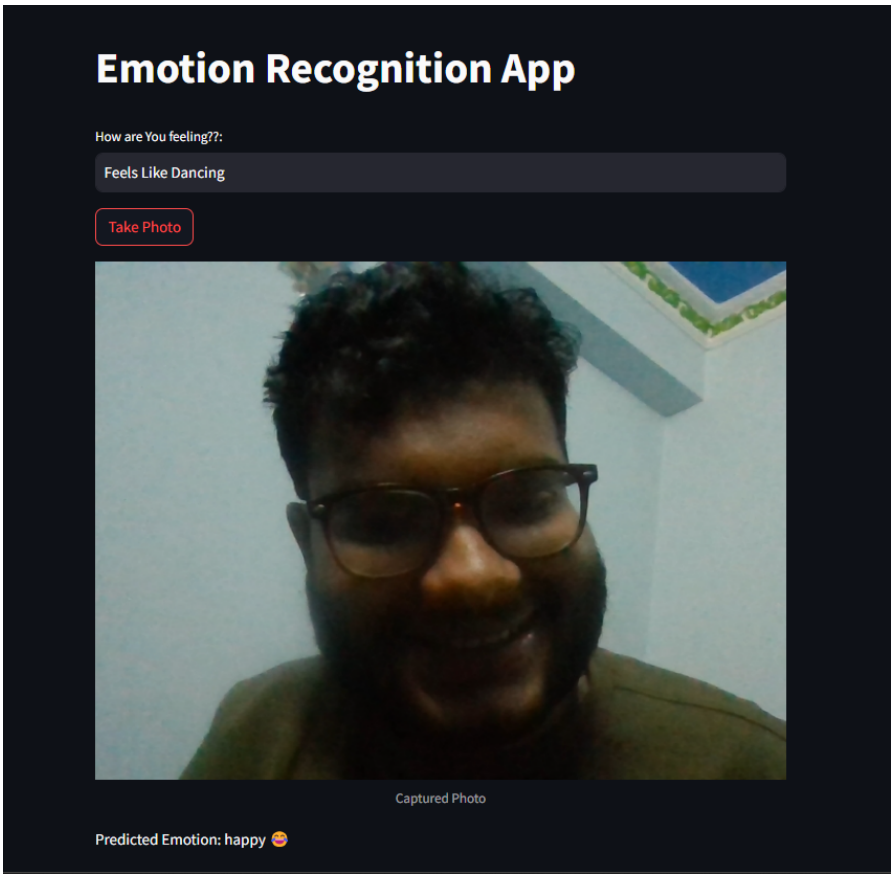

Figure A-2: True Prediction by the System: Output 2

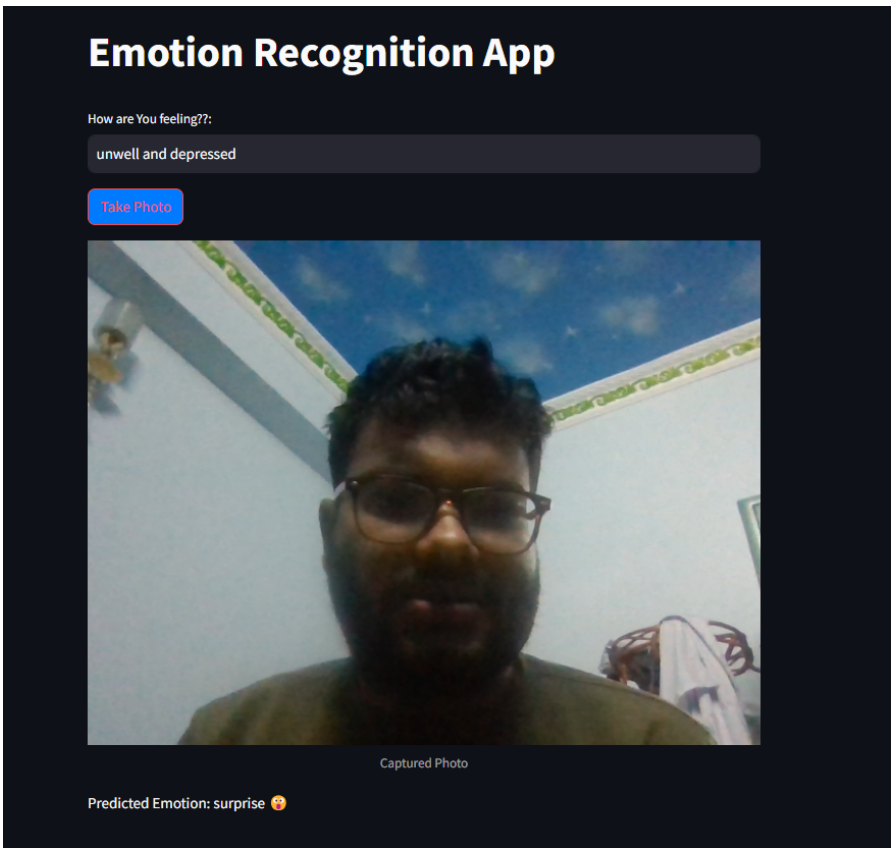

Figure A-3: False Prediction by the System: Output 1

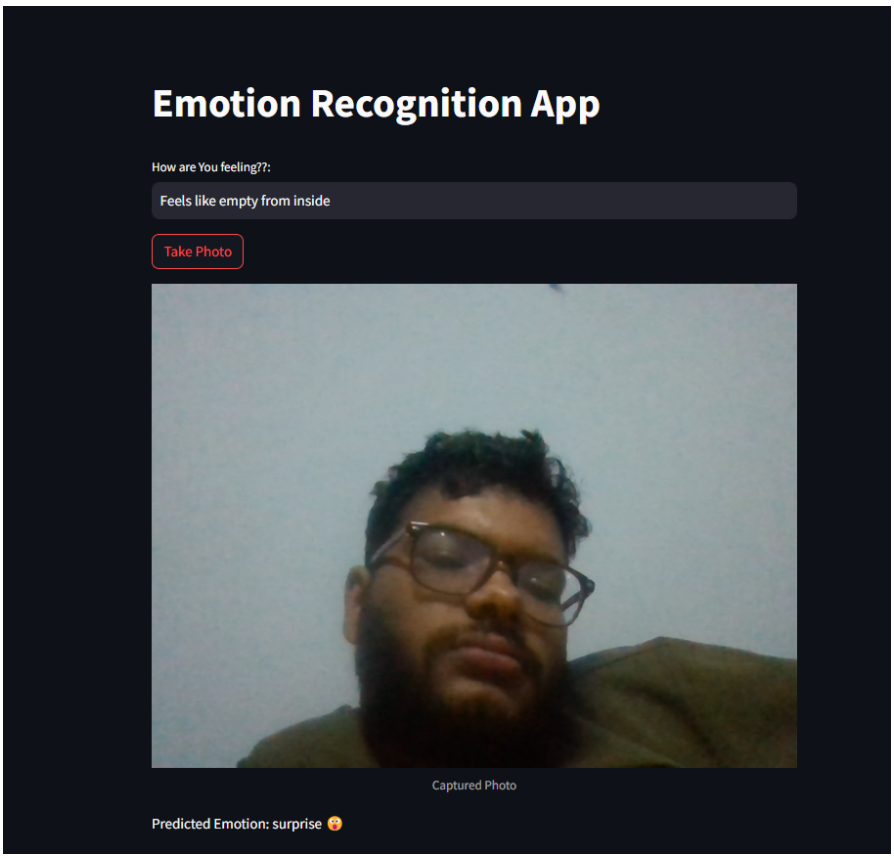

Figure A-4: False Prediction by the System: Output 2

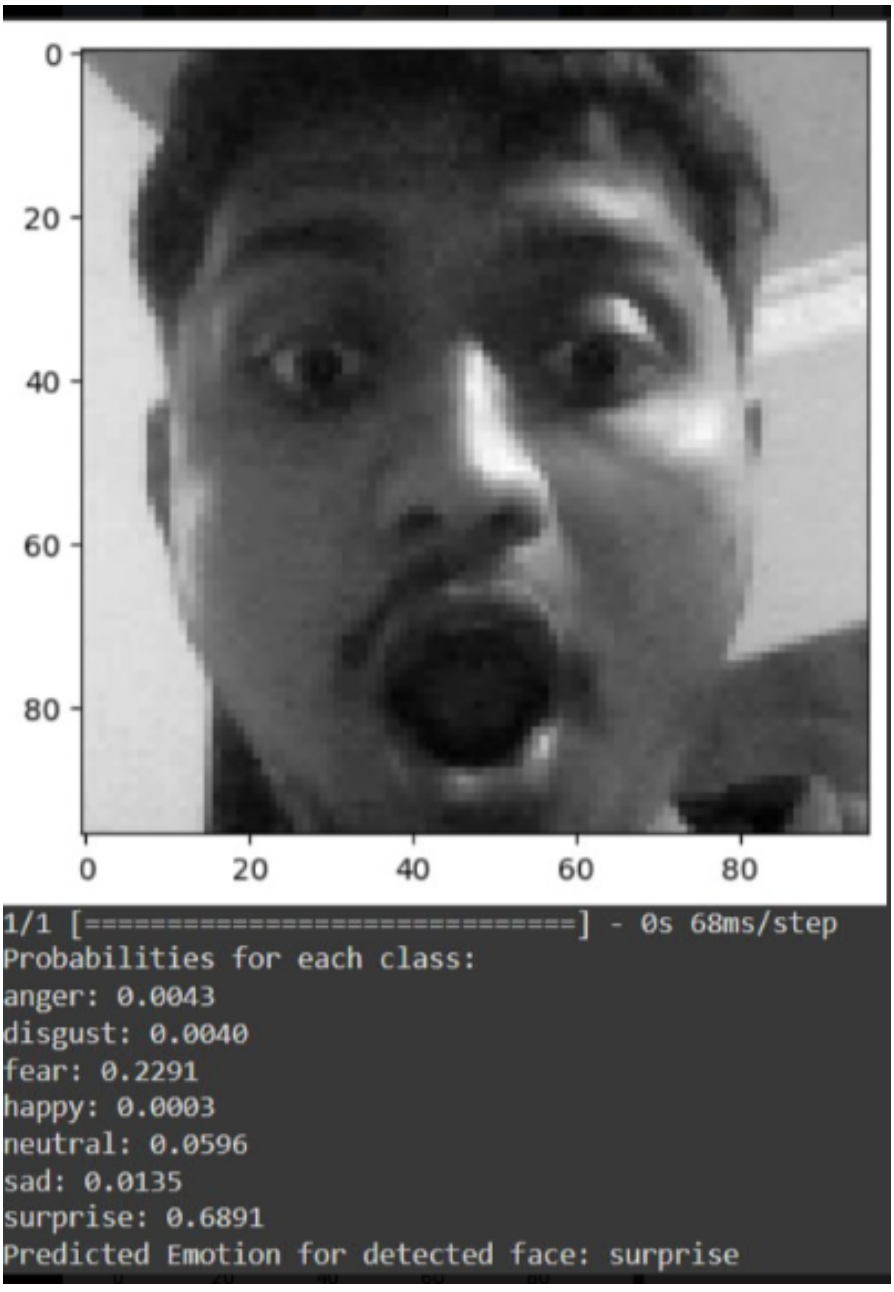

Figure A-5: True Prediction by the Custom ConvNet Model: Output 1

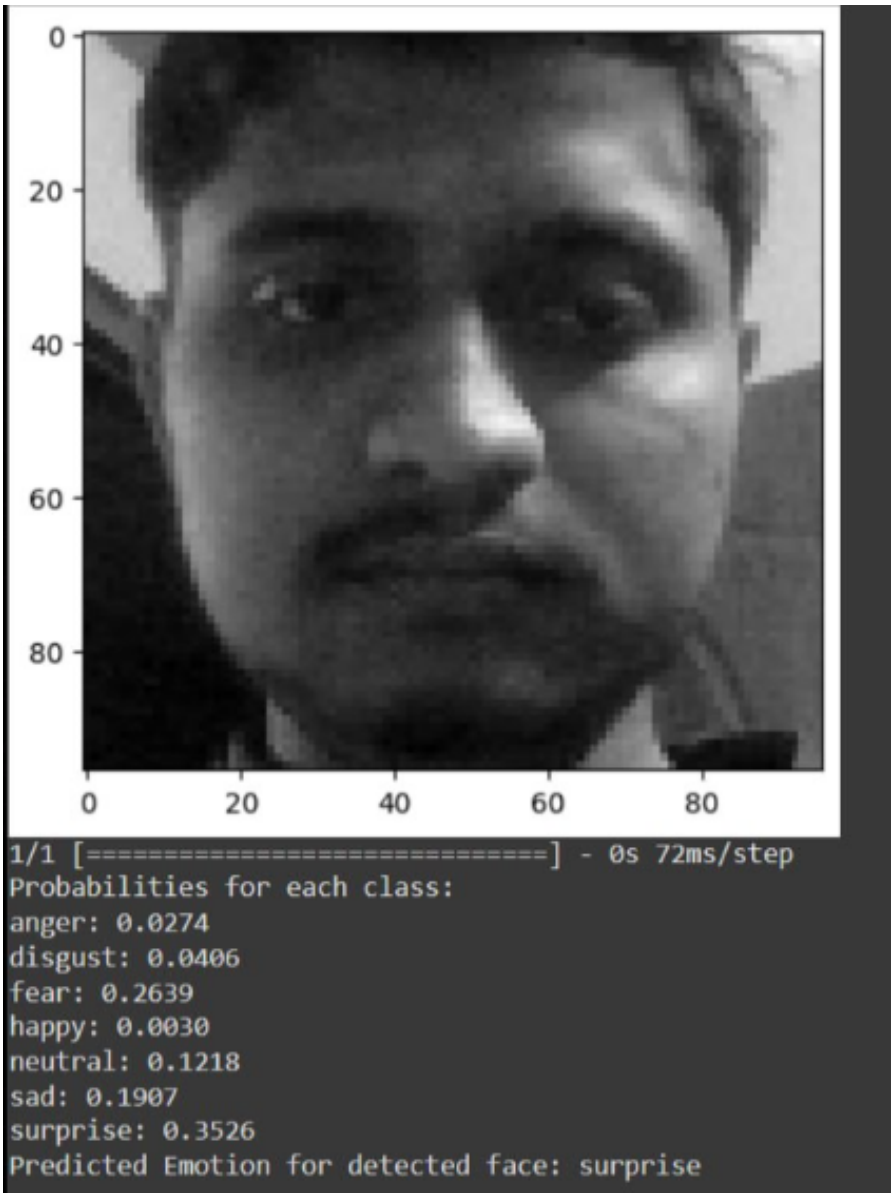

Figure A-6: False Prediction by the Custom ConvNet Model: Output 1

```
Enter sentence: Hello He is Hari
Predicted Emotion: neutral
Probability of Emotions:
anger: 0.023580297636128632
disgust: 0.0752008594297268
fear: 0.0355785035617441
happy: 0.08511323203074754
neutral: 0.5953414457459522
sadness: 0.12456922273235299
surprise: 0.06061643886334772
```

Figure A-7: True Prediction by the Fined Tuned BERT Model: Output 1

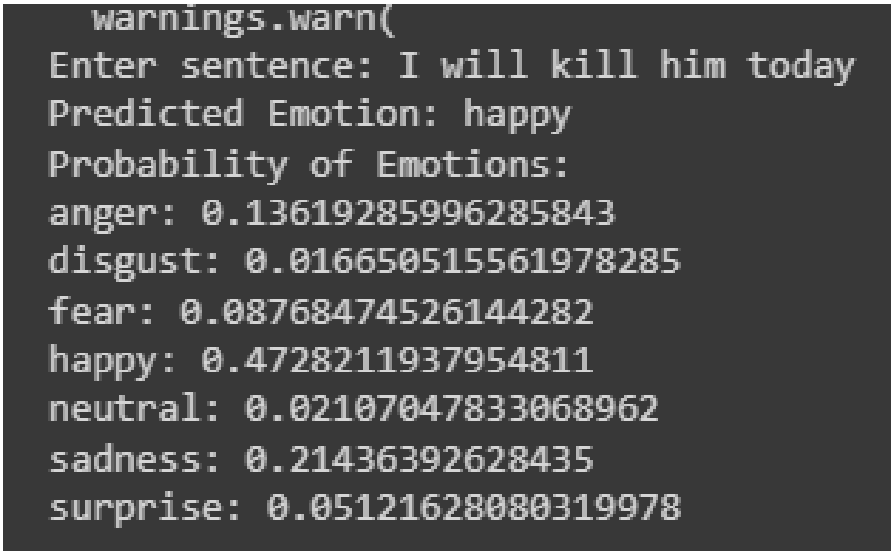

Figure A-8: False Prediction by the Fined Tuned BERT Model: Output 1