FACULTY OF ENGINEERING AND ARCHITECTURE

REAL-TIME EMOTION DETECTION **(FACE - VOICE)**

ARTIFICIAL INTELLIGENCE APPLICATIONS

ELİF NUR ASLIHAN CELEPOĞLU

1904010023

FINAL PROJECT REPORT

Dr. ABDULKADİR KAYIKLI

CONTENTS

1. Introduction

The importance of emotion recognition from facial expressions and voice is rapidly growing, with expanding applications in human-computer interaction, security, and entertainment. Numerous techniques have been employed to enhance the precision and efficiency of emotion recognition systems, leveraging Python libraries and various deep learning frameworks.

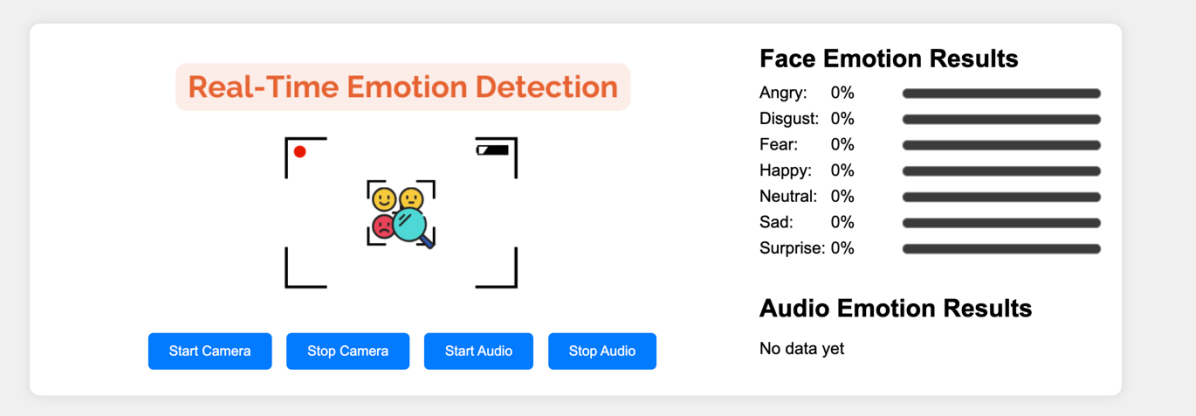
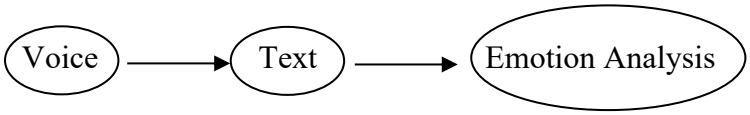


Figure1 interface view of the application

2.Emotion Recognition from Voice

It involves the process of recognizing emotions from voice. Voice recording is made, the recording is converted into text, and then sentiment analysis is performed on the text.



2.1 Information Dataset : (RoBERTa Dataset)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8187 entries, 0 to 8186
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Id                      8187 non-null   int64
 1   ProductId               8187 non-null   object
 2   UserId                  8187 non-null   object
 3   ProfileName             8187 non-null   object
 4   HelpfulnessNumerator    8187 non-null   int64
 5   HelpfulnessDenominator  8187 non-null   int64
 6   Score                   8187 non-null   int64
 7   Time                    8187 non-null   int64
 8   Summary                 8187 non-null   object
 9   Text                    8187 non-null   object
dtypes: int64(5), object(5)
memory usage: 639.7+ KB
```

Dataset link : https://github.com/yueyu1030/COSINE/tree/main/data

2.2 Steps Undertaken :

- Audio is recorded and saved in a .wav file.
- The audio recording is converted to text using the Hugging Face API.
- The resulting text is sent to the j-hartmann/emotion-english-ditilroberta-base model (.h5 file name) and sentiment analysis is performed on the text.
- Analysis results are returned as a probability value for each emotion.
- Emotion possibilities are printed on the screen.

Modeling

Masked Language Modeling (MLM) is used to train the language model. By masking some words or tokens within the text, the model learns to predict these hidden words, improving its understanding of language context for more accurate predictions. The method used here performs sentiment analysis from text, incorporating audio by converting it into text and then analyzing the sentiment.

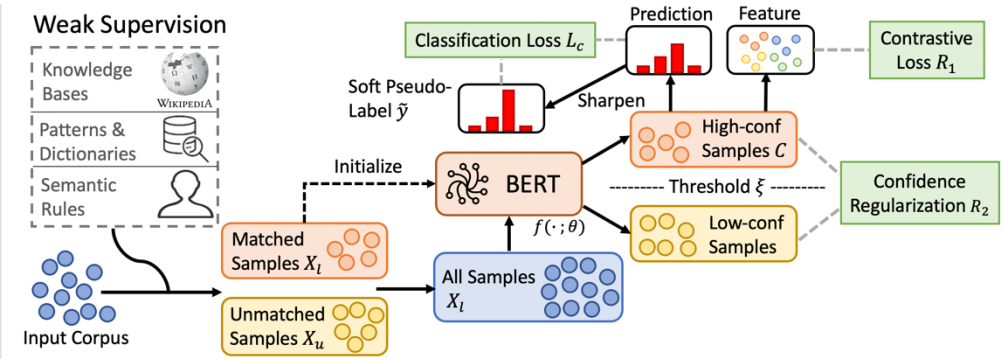
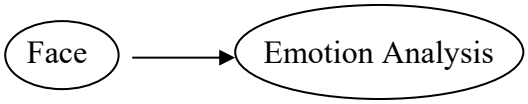


Figure2 diagram of MLM

3.Emotion Recognition from Face

It involves the process of recognizing emotions from facial images. Faces are detected in the images taken from the camera, the faces are cropped and then emotion analysis is performed on the cropped face images.



3.1 Information Dataset : (Fer2023 Dataset)

The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image.

The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

The training set consists of 28,709 examples and the public test set consists of 3,589 examples.

Dataset link : https://www.kaggle.com/datasets/msambare/fer2013

Modeling

The model is based on the Vision Transformer (ViT) architecture, which has gained popularity in image processing. Originally successful in language models, the Transformer architecture has been adapted for image processing tasks.

Model Features

- ViT utilizes Transformer blocks instead of convolutional neural networks (CNNs) for image processing tasks.
- The model is trained for facial expression recognition tasks using a dataset of facial photographs representing different emotional expressions.
- It is capable of recognizing and classifying emotional expressions.

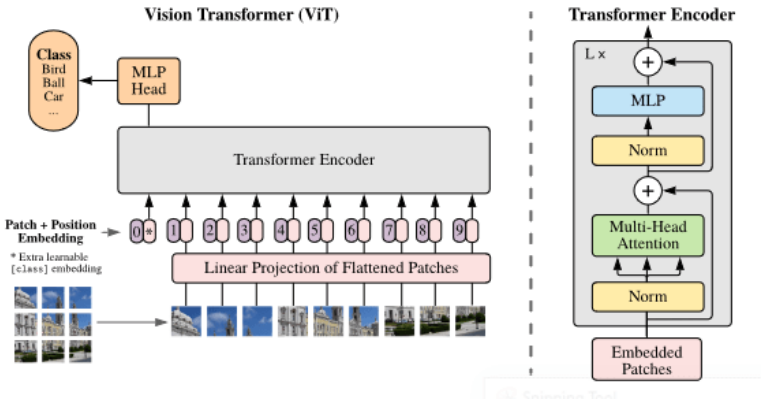


Figure3 diagram of VIT

3.2 Steps Undertaken

- The image is taken from the camera and an image frame is obtained using OpenCV.
- Using the MTCNN model, faces in the image are detected and cropped.
- The cropped face image is sent to the ViT model and emotion analysis is performed.
- The model produces a probability score for each emotion class.
- Probability scores are returned as a dictionary.
- The cropped facial image and emotion possibilities are sent to a drawing function.
- The draw function displays the facial image and a bar graph of emotion probabilities.
- The result image is saved in the "static/results.png" file.

4.Multimodel Emotion Recognition from (Face – Voice)

At this stage, a multi-model emotion recognition application was created by combining the emotion recognition results obtained from voice and facial image data. A web application has been developed using the Flask web framework.

4.1 Steps Undertaken:

- First, voice.py and face.py files are included in the project. These files perform emotion recognition from voice and facial image data, respectively.
- Then, the obtained emotion recognition results were defined as global variables and used within the functions of the Flask application in the app.py file.
- HTML templates have been prepared to create the user interface. These templates define the visual design and interaction elements of the web application.
- To improve user experience, buttons have been added to process audio and facial image data. These buttons allow the user to easily start emotion recognition processes and view the results.

4.2 Result

At the same time, sentiment analysis is being carried out on the website. You can only control it by pressing the start and stop buttons.

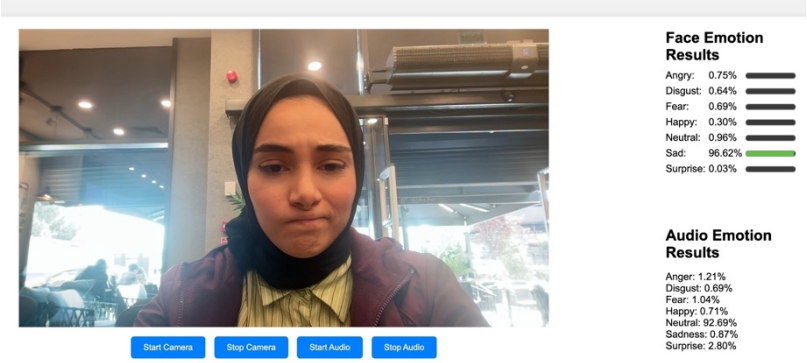


Figure 4 Result of the application

References

Kaggle. (n.d.). FER2013 Dataset. Retrieved from https://www.kaggle.com/datasets/msambare/fer2013

Arriaga, P., & Poggio, T. (2017). Synthesized Classifier Performance Comparison for Facial Expression Recognition. arXiv preprint arXiv:1708.03985.

Lim, B. C., & H., R. L. (2018). A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors, 18(2), 401. https://doi.org/10.3390/s18020401

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Zenodo. (n.d.). Linguistic Data Consortium. Retrieved from https://zenodo.org/records/1188976#.Y-JJJ-zMKUk

Hugging Face. (n.d.). Emotion-English-DistilRoBERTa-Base. Retrieved from https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Conneau, A., & XNLI. (2018). Cross-lingual Language Understanding through XNLI. arXiv preprint arXiv:1809.05053.

Hugging Face. (n.d.). How to train your own NLP model. Retrieved from https://huggingface.co/blog/how-to-train

Kar, A., Ramakrishnan, A., Srikumar, V., & Bhattacharyya, P. (2020). Emotion recognition from text using bidirectional LSTM model with BERT embeddings. arXiv preprint arXiv:2010.07835