Multimodal Emotion Recognition Using Deep Learning and Fusion Techniques

Project Advisor: Adnan Yazici
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
adnan.yazici@nu.edu.kz

Project Co-Advisor: Enver Ever
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
enver.ever@nu.edu.kz

Ainur Khamitova
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
ainur.khamitova@nu.edu.kz

Temirlan Nurmakhan
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
temirlan.nurmakhan@nu.edu.kz

Zhanna Mukhametsharip
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
zhanna.mukhametsharip@nu.edu.kz

Zhazira Kabdrakhmetova
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
zhazira.kabdrakhmetova@nu.edu.kz

I. Executive Summary

Emotion recognition plays a crucial role in human-computer interaction, significantly influencing the advancement of virtual assistants, mental health diagnosis tools, and customer experience analysis systems. Our senior project aims to develop an advanced multimodal emotion recognition (MER) model using modern deep learning techniques and fusion methods.

Most traditional emotion recognition models rely on a single modality for decision-making, such as facial expressions or text. However, this approach can be limited in capturing the complexity of human emotions. To overcome this limitation, we will integrate multiple input types to create a more comprehensive model, reducing misclassifications and improving overall system performance.

Our system includes an emotion recognition model and a user interface for interaction. The web application will serve as the interface, allowing users to upload video materials of a specified duration. The application extracts audio, video, and text from the uploaded video and feeds them into different deep-learning models customized for each modality. The outputs, representing probabilities for various emotion classes (e.g., "happy," "sad," "fearful," "surprised," "angry," "disgusted," and "neutral"), will be combined using fusion techniques for enhanced accuracy. The web app then presents visual representations of the emotions through graphs and descriptions for user interpretation.

II. Introduction

Nowadays, research in affective computing has explored two fundamental approaches to exploring human emotions: unimodal and multimodal recognition. Unimodal systems analyze emotions using just one modality, such as facial expressions from a video, voice tone and speech, or textual context. Nevertheless, studies have revealed that human emotions are too complex to be fully captured by unimodal methods [1]. In contrast, multimodal emotion recognition models, which combine data from multiple inputs, such as text, visual, and audio provide a more comprehensive portrayal of emotional states. By integrating diverse modalities, these models offer a better understanding of emotions, breaking the limitations of an unimodal approach. In addition, there is no comprehensive user interface that allows users to interact and test the model. That is why, the main purpose of our project was to develop a multimodal emotion recognition web application. The motivation behind our project comes from this need to enhance the efficacy and accuracy of emotion recognition systems and provide a comfortable UI that helps to utilize the model in real life.

Our solution is concentrated on the development of a comprehensive multimodal emotion recognition model, augmented by modern deep learning techniques and fusion methods. The system applies different methods for processing audio, video, and textual inputs extracted from user-uploaded videos. These inputs are fed into specialized deep-learning models customized for each modality, yielding probabilistic outputs of various emotion classes. The outputs are fused using the late fusion technique and manipulated through a user-friendly web application that enables convenient interaction with output emotions for user interpretation.

The report is structured to provide a comprehensive overview of our project, with a detailed description of the various stages of development and the methodology employed. The report begins with an overview of previous research in the area of affective computing, followed by a detailed description of the project approach, which includes explanations of model and web application development. Next, project execution reports for the fall and spring semesters are outlined, and the challenges faced are discussed, along with their solutions. Finally, a thorough evaluation of the utilized models and fusion methods is provided, and conclusions are drawn with suggestions for future improvement.

III. Background and Related Work

A. Unimodal emotion recognition

Unimodal emotion recognition using voice, facial expressions or text features is recognized for its simplicity and low computational cost in classification:

1) For image classification task, Convolutional Neural Networks (CNN) is robust and effective in performing high-level feature extraction Deep Learning architecture. In a study [4], a face recognition algorithm combined features from convolutional neural networks (CNNs) and the bag-of-visual-words (BOVW) model for handcrafted features, achieving high accuracy. Similarly, Agrawal and Mittal's research investigated CNN parameters' impact on emotion classification accuracy and proposed two CNN architectures with an accuracy of 65% on FER-2013 dataset [5].
2) For speech emotion recognition, Issa et al. introduced a CNN-based approach for speech emotion recognition by using convolutional layers followed by fully connected layers architecture for classification [6]. The model was developed and evaluated on the EmoDB, RAVDESS and IEMOCAP datasets reaching higher accuracy compared to state-of-the-art methods. In work [7], novel neural network architecture, SENN (Semantic-Emotion Neural Network), was proposed to capture emotional connections between words using Bi-LSTM and CNN.
3) For text emotion recognition, Pre-trained transformer models like BERT, RoBERTa, DistilBERT, and XLNet are increasingly used. In a study [8], these models were fine-tuned on the ISEAR dataset, achieving accuracy rates of 74.31%, 72.99%, 70.09% and 66.93% for RoBERTa, XLNet, BERT and DistilBERT, respectively.

However, unimodal emotion recognition has limitations in capturing specific user emotions in real-time. Combining multiple modal features, as suggested by Wei et al. and Zhang et al., provides a more comprehensive approach that surpasses unimodal recognition [14][15]. In the human-computer interaction research area, there's a focus on multimodal emotion recognition, which involves understanding connections within and between different modalities. Machine learning algorithms such as Support Vector Machine (SVM), Hidden Markov Model (HMM), and Gaussian Mixture Model, along with deep learning techniques, are applied to recognize emotions across various data modalities. Deep learning methods excel in feature extraction and are increasingly adopted for multimodal emotion recognition.

B. Multimodal emotion recognition

Human communication involves various channels beyond facial expressions; it includes nonverbal cues, including gestures, as well as verbal cues through speech, pitch, volume, and the flow of words in written text. Additionally, physiological signals like EEG, heart rate, blood pressure, and pulse contribute to emotional expression. The diversity in emotional communication suggests the possibility of enhancing machine emotion recognition by integrating a wide range of cues reflecting the human perceptual processes [16].To enhance emotion recognition accuracy, a proposed solution incorporates a deep learning-based multimodal fusion technique, originally introduced by Duc et al. (1997) [20].

Research conducted by Issa et al. (2011), Kahou et al. (2014), and Mittal et al. (2017) explored emotion extraction from speech alone, a multimodal approach integrating video, and combined facial, text, and speech cues, respectively [6], [19], [11]. Experimental findings demonstrate that the recognition performance of individual modalities is inferior to that of multimodal approaches. Mittal et al. (2017) introduced a robust multiplicative modality fusion method that effectively handles sensor noise [11]. Another approach proposed by Dai et al. (2021) integrates noisy input data replacement and feature-level fusion into advanced emotion recognition frameworks from audio, video, and text. This method emphasizes multi-class classification to capture the nuanced nature of human emotions [21]. In research [9], a multimodal emotion detection model using CNN-LSTM, which combines audio and visual information at the model level was presented. This model achieved an impressive accuracy of 90.06% on the MELD dataset. Late fusion methodologies, used by Pandeya et al. (2021), Kumar et al. (2022), and Ding et al. (2022), involve building separate models for each modality and integrating their results at the prediction level using techniques like averaging, weighted sum, or deep neural networks [22], [23], [25]. Research [22] presents two model variants for different conditions, while [25] introduces a novel deep neural network that combines audio, video, and text modalities, showcasing effectiveness across various environments.

In a recent study by Kumar et al. (2023), the challenges faced by deep learning-based multimodal emotion recognition systems were addressed, particularly focusing on the lack of interpretability and labeled datasets reflecting real-life multimodal scenarios [24]. To tackle these issues, ParallelNet was introduced, employing two networks, N1 and N2, for hybrid fusion of speech and image modalities. Spinal fusion techniques were utilized, combining intermediate and late fusion methods. Through careful architecture design and ablation studies, N1 and N2 were optimized to increase performance while maintaining interpretability. The proposed system achieved a reasonable accuracy of 83.29% on the IIT-R SIER dataset, with an emphasis on improving interpretability through feature interpretation techniques.

Milon Islam et al. utilized a combination of DSCNN architecture, Bi-LSTM technique, soft attention technique, GAP layer, and an emotion classification block [26]. For visual features, they employed a Depthwise Separable Convolution Neural Network (DSCNN) achieving 79.77%, and for physiological data (including Trapezius Electromyogram (tEMG), ECG, and SCL), they used Bi-directional Long Short-Term Memory (Bi-LSTM). Fusion was accomplished using a soft attention method, with Global Average Pooling used to concatenate features. Face detection methods included Histogram of Oriented Gradients and a CNN-based face detector (ResNet-

34) with accuracies of 95.30% and 96.04%, respectively. They tested their approach on the Bio Vid Emo DB multimodal dataset, consisting of five different emotions: Amusement, Anger, Disgust, Fear, and Sadness.

Ayata et al.[27] applied 2D CNN for verbal data processing and 3D CNN for video processing, using the DEAP dataset for training and evaluation. They utilized the minimum redundancy maximum correlation method and achieved accuracies of 73.08% and 72.18% for arousal and valence emotions, respectively. Testing on the eNTERFACE05 and BAUM-2 datasets resulted in accuracies of 87.02% and 56.12% for three types of emotions: happy, pain, and normal.

Barkur et al. [28] employed attention WaveNet and manual feature extraction from two parallel networks, achieving cumulative accuracies of 58%, 44% and 68% on the EMO-DB, CREMA-D and SAVEE datasets, respectively. They used a CNN-based face detection method with Local Binary Convolution (LBC). Some approaches presented by research papers [9] and [12 include CNN, CNN+RNN and CNN+LSTM to distinguish between different emotion labels and sentiments. Due to the varying data distributions and formats in different modalities (audio, video, text) implementing fusion mechanisms can be challenging. However, integrating models trained on various modalities presents challenges due to differences in data distributions and formats across modalities (audio, video, text). This can lead to sensitivity and generalization gaps among the combined models, potentially causing decision bias towards the most accurate model. To address this, we intended to explore early and late fusion approaches.

The process of emotion recognition encompasses several stages, including data collection, preprocessing, feature extraction, and classification [17]. To address the complexity of emotional signals, multimodal fusion has emerged as a promising approach, leveraging features from multiple modalities. Multimodal emotion recognition leads to increased computational demands and storage requirements [18]. The disadvantage is that high-dimensional data and may not effectively capture the intrinsic characteristics of each modality, leading to issues such as information redundancy and computational inefficiency.

In line with Milon Islam et al.'s work, our project uses CNN architecture for facial emotion recognition tasks, with a dropout to prevent overfitting and ReLU and softmax activation functions. Facial landmarks were identified using OpenCV. We applied learning rate, batch size, optimizer, and loss function settings, and incorporated early stopping. For audio classification as Barkur et al. work presented, the model was trained on three datasets. For test classification pre-trained transformer model RoBERTa was used and trained on GoEmotions dataset, considering significant performance presented in study [8]. To concatenate features at the decision level, a late fusion architecture was proposed for improved accuracy. The late fusion method outperformed early fusion, particularly when the three modalities proposed different results. Evaluation metrics such as Accuracy, Precision, Recall, and F1-Score were employed, with architectures implemented using TensorFlow Keras, Python language on Google Colab.

IV. Project Approach

A. System Description

The proposed system consists of a web interface that interacts with a multimodal emotion recognition model utilizing text, image, and audio modalities to classify video segments into one of seven emotions. As shown in Fig. 1., the sequence diagram outlines four main components: User actions, Web interface, Server (back-end), and Multimodal Emotion Recognition (MER) tool. The following interactions are possible between these components:

- **The User** interacts with the system through actions such as login, registration, and uploading videos.
- **The Web interface** serves as the front-end, providing visual tools and interfaces for user interaction.
- **The Server** handles user authentication, database storage, and communication between users and the MER tool.
- **The MER** analyzes uploaded videos using image, audio, and text modalities, providing emotion recognition data to the web interface.

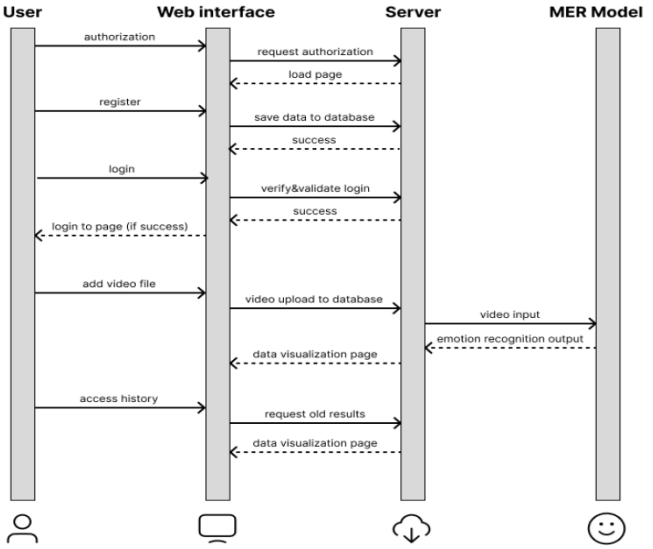

Fig. 1. Sequence diagram of the proposed system

The general model workflow is shown in Fig. 2. As displayed, the proposed model extracts input data from the input video. Then they are passed to corresponding models, where prediction softmax for each modality is obtained. These predictions are then passed into the fusion where it is further processed and given as an output.

The following third-party tools were utilized during project development:

- **Google Colab Pro:** For quicker model training, we utilized Colab Pro subscription which allowed quicker execution of the training. Given the fact that datasets included thousands of samples and were used to train complex models, using general software would take too long. Unfortunately, the general version of Colab contin-

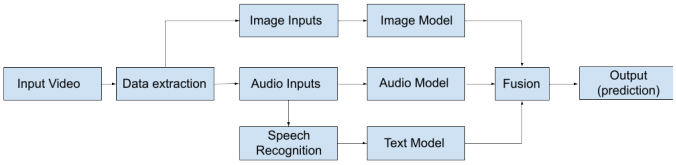

Fig. 2. General workflow of the model

uously crashed due to RAM and GPU usage limitations for general users.

- **GitHub and Google Drive:** These platforms were used to access datasets through the Google Colab and Python interfaces. Uploading a dataset each runtime would be time-consuming which is why they were directly accessed through GitHub links or a local copy uploaded to Google Drive.
- **Visual Studio 2022 and MongoDB Compass:** Visual Studio 2022 is an integrated development environment (IDE) by Microsoft, which provides a comprehensive set of tools and plugins for web development, such as code editing, debugging, and version control. MongoDB Compass is the official graphical user interface (GUI) for MongoDB, a popular NoSQL database management system. It provides a convenient visual representation of their MongoDB databases, where collections can be explored, queried, and manipulated without the need for command-line interactions.

B. Model Development

1) Textual Emotion Recognition Model: For the development of the text model state-of-art transformers provided by HuggingFace were obtained. Initially, we used the MELD dataset's textual data [26], which was saved in the CSV file. It included "Utterance" and "Emotion" fields among others, and these two were used to fine-tune the pre-trained BERT model. The Ktrain library was used for the training purpose. During the model training phase, parameters like 2e-05 learning rate and 3 epochs were used.

State-of-art F1-score is 65.8% [27] and the results from the BERT model were 62% weighted F1-score. As we can see in Table III and IV from Appendices, it is a generally good result, however, the accuracies of underrepresented classes, like Fear, Disgust, and Sadness were below 40%, which was due to the imbalanced nature of the given MELD dataset. Randomly oversampling these classes did not yield better results, which is why it was concluded that it might be better to look for another dataset with better-quality textual data.

For this reason, we found another dataset, GoEmotions, which was curated by Google and has 27 emotional classes. For the sake of adhering to our initial 7-class model due to training audio and visual modalities in that manner, we selected a subset of a large GoEmotions dataset. This included 7 universal emotion classes. For this task, we utilized another transformer model, namely RoBERTa from the TFRobertaForSequenceClassification class (See Fig. 3). The model was trained on 12620 data samples with 16 batches for 3 epochs. Tokenizer instance of RobertaTokenizerFast class from 'roberta-base' pre-trained model was used to preprocess the text. The model itself was compiled using the Adam optimizer with a learning rate 5e-5, and Sparse Categorical Crossentropy was set to evaluate the loss.

```
Model: "tf_roberta_for_sequence_classification"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 roberta (TFRobertaMainLaye  multiple                  124055040
 r)

 classifier (TFRobertaClass  multiple                  595975
 ificationHead)

=================================================================
Total params: 124651015 (475.51 MB)
Trainable params: 124651015 (475.51 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig. 3. Structure of the RoBERTa model

2) Speech Emotion Recognition Model: All audio used in training and validation is preprocessed to remove noise from the background. Input audio was augmented with NoiseAddition, Shifting, Pitching, and Stretching methods to generate more data for training and evaluation.

In the development of speech emotion recognition, the implementation of features extracted from input audio files and used for training of the model was complete. For this purpose Mel-Frequency Cepstral Coefficients (MFCCs) were used which are crucial in speech and audio processing, providing a compact representation of the power spectrum of sound on a perceptually relevant scale (See Fig. 4).

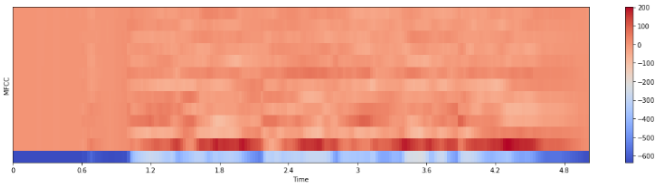

Fig. 4. Structure of the speech CNN model

As a model, a Convolutional Neural Network (CNN) was developed (See Fig. 5). The model architecture includes a series of Convolutional layers followed by MaxPooling1D layers, employing 32, 64, 128, 128, and 256 filters with a filter size of 6x6 and strides of 1. MaxPooling layers have a stride of 2. Towards the end, a Dropout layer is incorporated, randomly discarding 2% of the data, followed by a fully connected (FC) layer with 32 neurons with a 'Relu' activation function. The final layer utilizes softmax activation for the 7 output classes. This model is lightweight, containing only 359783 parameters, all of which are trainable, yet it achieves commendable accuracy. For optimization, the model utilizes the Adam optimizer and employs categorical cross entropy

as the loss function, with accuracy serving as the primary evaluation metric.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 162, 128)	768
max_pooling1d (MaxPooling1D)	(None, 81, 128)	0
conv1d_1 (Conv1D)	(None, 81, 128)	82,048
max_pooling1d_1 (MaxPooling1D)	(None, 41, 128)	0
conv1d_2 (Conv1D)	(None, 41, 32)	20,512
max_pooling1d_2 (MaxPooling1D)	(None, 21, 32)	0
dropout (Dropout)	(None, 21, 32)	0
conv1d_3 (Conv1D)	(None, 21, 32)	5,152
max_pooling1d_3 (MaxPooling1D)	(None, 11, 32)	0
lstm (LSTM)	(None, 11, 128)	82,432
batch_normalization (BatchNormalization)	(None, 11, 128)	512
lstm_1 (LSTM)	(None, 64)	49,408
batch_normalization_1 (BatchNormalization)	(None, 64)	256
flatten (Flatten)	(None, 64)	0
dense (Dense)	(None, 128)	8,320
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 7)	903

Total params: 750,167 (2.86 MB)
Trainable params: 249,927 (976.28 KB)
Non-trainable params: 384 (1.50 KB)
Optimizer params: 499,856 (1.91 MB)

Fig. 5. Structure of the speech CNN model

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv2d_8 (Conv2D)	(None, 48, 48, 20)	200
conv2d_9 (Conv2D)	(None, 48, 48, 30)	5,430
max_pooling2d_3 (MaxPooling2D)	(None, 24, 24, 30)	0
batch_normalization_2 (BatchNormalization)	(None, 24, 24, 30)	120
dropout_3 (Dropout)	(None, 24, 24, 30)	0
conv2d_10 (Conv2D)	(None, 24, 24, 40)	10,840
conv2d_11 (Conv2D)	(None, 24, 24, 50)	18,050
max_pooling2d_4 (MaxPooling2D)	(None, 12, 12, 50)	0
batch_normalization_3 (BatchNormalization)	(None, 12, 12, 50)	200
dropout_4 (Dropout)	(None, 12, 12, 50)	0
conv2d_12 (Conv2D)	(None, 12, 12, 60)	27,060
conv2d_13 (Conv2D)	(None, 12, 12, 70)	37,870
max_pooling2d_5 (MaxPooling2D)	(None, 6, 6, 70)	0
dropout_5 (Dropout)	(None, 6, 6, 70)	0
conv2d_14 (Conv2D)	(None, 6, 6, 80)	50,480
conv2d_15 (Conv2D)	(None, 6, 6, 90)	64,890
flatten_1 (Flatten)	(None, 3240)	0
dense_3 (Dense)	(None, 1000)	3,241,000
dense_4 (Dense)	(None, 512)	512,512
dense_5 (Dense)	(None, 7)	3,591

Total params: 11,916,411 (45.46 MB)
Trainable params: 3,972,083 (15.15 MB)
Non-trainable params: 160 (640.00 B)
Optimizer params: 7,944,168 (30.30 MB)

Fig. 6. Structure of the image CNN model

3) Facial Emotion Recognition Model: In order to get the input for facial emotion recognition model, first the video was divided into frames and each 10th frame was analyzed by the model. Using dlib, OpenCV, and imutils library, the faces of the speaker were detected and facial landmark indices were defined. As described in Fig. 6., the CNN model was developed to predict facial emotions.

CNN-based model inspired by the VGGFace architecture was used consisting of 16 convolutional layers grouped into blocks, followed by fully connected layers for classification was used. The convolutional layers have varying filter sizes: 64, 128, 256, and 512 filters. The input shape for the model is (224, 224, 3), representing the dimensions of the input images. The model is compiled using the Adam optimizer with a learning rate of 0.001. The loss function is Sparse Categorical Crossentropy, suitable for multi-class classification tasks. Training is performed for 100 epochs with a batch size of 32. Early stopping is implemented with a patience of 8 to prevent overfitting. A pre-trained cascade classifier is used for face detection, implemented using the Haar cascade classifier. Detected faces are cropped from the input frames for emotion recognition. The average emotion predictions are computed from the predictions obtained for each frame.

4) Fusion Implementation:

1) *Early fusion*: The baseline model for early fusion includes ConvLSTM2D, Conv2D, Conv1D, MaxPooling2D, MaxPooling1D, GlobalMaxPooling1D, Embedding, Flatten, and Dense layers. The model produces two outputs: one for emotion classification with 7 classes and another for a different task with 3 classes. Since our task evaluates only emotions the early fusion approach concatenates GlobalMaxPooling, Dense and Flatten layers for text, audio and image classification models and returns a single output for emotion classification for MELD dataset. Training involves optimizing model parameters to minimize a specified loss function, such as SparseCategoricalCrossentropy. Techniques like early stopping was employed to prevent overfitting during the training process. The performance was 47% since we excluded sentiments and the model was heavy, and we considered late fusion approach.
2) *Late fusion with weights*: The late fusion approach takes three lists of emotion prediction results from different modalities (audio, image and text) and computes the result based on the following: Each emotion prediction result in the input lists is multiplied by a weight to adjust its contribution to the final fusion result. The weights are applied to each modality's results based on preset conditions related to the standard deviations of the results. If the standard deviation of a modality's results exceeds a certain threshold, a higher weight is assigned to that modality. Conversely, if the standard deviation is below a minimum threshold, a lower weight is assigned. The weighted results from each modality are combined to form the fused result.
The model was also tested for MELD dataset, and in comparision with early fusion succeeded in identifying distinctive emotions like 'happy', 'sad', 'surprise' and 'angry' resulting in f1-scores 0.41, 0.40, 0.42, 0.35 for a balanced testing dataset but showed lower performance on the remaining three emotions. The performance dis-

crepancies can be attributed to several factors within the dataset. Incomplete multimodal data instances in certain videos, the presence of multiple faces in frames, and the inherently facetious nature of the dataset, derived from comedic series, all contribute to inconsistencies in the results for 'fear', 'disgust' and 'neutral' emotion labels.

C. Web Application Development

The development of web applications can be separated into 3 main sections - front-end development, back-end development, and model integration into the backend. A full list of used third-party tools and libraries is listed in Table VII - VIII in the Appendix section. In this section, the most important steps of the development will be described along with the reasons for the choice of development tools.

1) Front-End Implementation: The first step in the development of the application was choosing the framework for front-end development. React.js was chosen due to its popularity, extensive ecosystem, and component-based architecture. React.js enables efficient development of dynamic user interfaces with its virtual DOM, JSX syntax, and state management capabilities. Additionally, its integration with libraries like React Router for declarative routing and React-bootstrap for pre-designed UI components simplified the development process. React Redux was used to manage the centralized application state, which could be accessed throughout the pages. Finally, the axios library was used to make asynchronous HTTP requests to the backend. The figures containing the screenshot of pages can be seen in Fig.15 - 22 in Appendices.

2) Back-End Implementation: We choose Flask as the web framework for back-end development because of its adaptability, simplicity, and ease of use. Flask is suitable for creating RESTful APIs and managing HTTP requests because of its lightweight and simple design. Furthermore, Flask's wide ecosystem of extensions—which includes Flask-SQLAlchemy for database connectivity and Flask-JWT-Extended for JSON Web Token authentication—offers reliable solutions for a variety of backend services, such as implementation of user authentication and authorization. For the database hosting, MongoDB was chosen, as it is a popular open-source NoSQL database famous for its scalability and ease of use. As depicted in Fig. 7 , the classes that describe collections were implemented in Flask. User collection stores documents of different users registered to the system. Video collection holds paths and titles of the uploaded video for a particular user. The results of the analyzed video are stored as documents in the Clip collection, which is used for history display.

3) Model Integration: We used TensorFlow and its Python API, Keras, to include machine learning models in the back-end. TensorFlow is a potent open-source framework that provides comprehensive support for deep learning methods for creating and training machine learning models. TensorFlow and Keras offer a complete model integration solution that makes it possible to deploy and run machine learning algorithms within the back-end quickly and effectively. Most model architectures were stored in a .keras file that stores the weights and structure of the individual model. Libraries, such as Librosa, OpenCV, and SpeechRecognition, were extensively used in the extraction and analysis of modality features. The resulting clips were combined using Moviepy and stored in the database for future access.

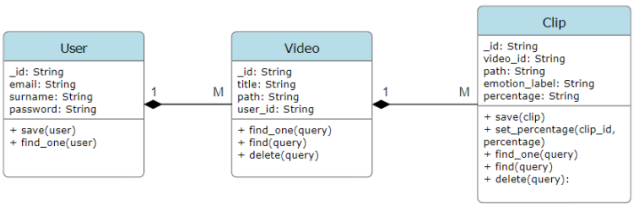

Fig. 7. UML for MongoDB database collections

V. Project Execution

Over the past two semesters, our senior project team has made significant progress in developing our multimodal emotion recognition model and web application. At the beginning of the fall semester, we allocated the tasks across periods of the semesters (See Fig. 23 in Appendix). During the Fall semester, we concentrated on the research and the development of facial and textual models, as well as initial architecture planning. This work continued in the Spring semester, where the main focus was to experiment with transformers, which was achieved through utilizing BERT, RoBERTa, and Vision Transformer throughout the experimentation period of the project creation. Another priority was to create an accessible web platform that combines all the models in its back end. This section provides a detailed description of steps taken in each semester.

A. Fall Semester

1) Initial Planning and Design Architecture: We began by conducting an extensive literature review on multimodal emotion recognition, exploring various deep learning techniques, fusion methods, and existing models. Based on our research, we outlined the architecture of our system, including the selection of deep learning models for each modality (audio, video, text), fusion techniques, and the design of the user interface.

2) Requirement Specification: After designing the basic architecture, we spent a noticeable effort on identifying functional, non-functional, and domain requirements both for the model and the web app. Further, the initial specification was modified after carefully reviewing the resource availability and adapting to our progress.

3) Development of the Facial Emotion Recognition Model: Initial FER model was based on a CNN approach and a variety of other Machine Learning principles. The FER-2013 dataset was utilized for this task. The CNN model demonstrated superior accuracy compared to alternative models such as SVM and LDA, prompting the team to concentrate on

exploring various configurations of CNN models further. At that time, our highest accuracy standed at 60.83%, while the best accuracies achieved on the FER2013 dataset typically hover around 72.5%.

4) Development of the Textual Emotion Recognition Model: In the previous semester, we utilized such libraries like FastText, as well as methodologies like Tf.Idf, SVC vectorizer, and MultinomialNB. These showed a maximum of 40% accuracy. The MELD dataset was utilized for this task. Hence, we decided to explore other methods of textual emotion recognition models.

B. Spring Semester

1) Audio Emotion Recognition Model: In this semester, we explored different ways of developing speech emotion recognition models. The various models, such as transformers and CNN models were trained on different datasets, such as Ravdess, TESS, and CREMA-D. After experimenting with hyperparameters and fine-tuning the models on different datasets, our CNN model showed the best result and achieved reasonable results.

2) Model Training and Optimization: In addition, we explored different ways to improve the models, developed during the Fall semesters. We tried different methods of transfer learning and hyperparameter tuning. In addition, for the text emotion recognition model, transformers, such as BERT and RoBERTa were investigated, and trained on the MELD and Goemotions dataset. As a result, the last model was chosen as our final text emotion recognition model.

3) User Interface Development: Simultaneously, we worked on developing the front end of our web application, focusing on creating an intuitive and interactive interface for users to upload videos, visualize emotional predictions, and interpret the results. We used frameworks like React for front-end development, ensuring cross-browser compatibility and responsiveness. In addition, APIs in the backend were developed using Flask, and the database for users and MER results was created.

4) Integration and Fusion: Once we had trained models for each modality, we integrated them into our system and implemented fusion techniques to combine their outputs. We explored late fusion as it was the least time-consuming method in development and integration. The performance of different methods for late fusion was evaluated and the best formula was chosen as a crucial part of the system.

5) User Testing: A self-made dataset was collected through questions provoking emotions from the person answering the questions (See Fig. 24 in Appendix). These questions were used to film 30-second - 1-minute videos for real-world testing cases. The videos were fed into the model and tested to measure the results and performance of the web app.

C. Challenges and Solutions

1) Schedule Change: There were some schedule changes because the work's focus was shifted towards utilizing transformers. This left the previously made models unused (such as CNN, etc.) Therefore, it took some time to adjust to the new goal of utilizing transformer models.

2) Data Availability: There were some problems with accessing certain datasets. The IEMOCAP dataset is only accessible through institutional email. It also weighs 16.5GB which made it difficult to access it at first. However, this issue was resolved later by choosing the subset for testing purposes. The CMU-MOSEI dataset was difficult to find on the Web which made it impossible to use it for training or testing.

3) Model Integration: Because models were previously written in Keras and Tensorflow, their integration was quite challenging due to version compatibility issues. Some models like ViT and other transformer models took time to be integrated into the previously established Tensorflow environment. The problem was solved by standardizing the versions of libraries and re-training the models to achieve up-to-date .keras files.

4) Performance Optimization: Due to the specifics of the audio model's output and softmax results from other models, it was challenging to integrate the models' results together in a late fusion stage. The final classification results were significantly shifted towards the audio model's results. This issue was resolved later by experimenting with the fusion and performance saw some improvements.

VI. Evaluation

In this project, we conducted different methods of evaluation. First each model was evaluated on different datasets and compared with state of art results (See Table I/). Moreover, different late fusion techniques were tested on datasets such as MELD and IEMOCAP. Lastly, the user interface was evaluated on the custom dataset and different information on the required resources were gathered.

A. Text Emotion Recognition Model

The Model was tested on the subsection of GoEmotions dataset. As shown in Fig. 8 - 9, the results we achieved a slightly better weighted F1 score of 65%. The main benefit was, though, in a better-distributed confusion matrix. All classes showed F1 scores above 40%, which is a big improvement compared to the model fine-tuned on the MELD dataset.

B. Speech Emotion Recognition Model

Train of model included 75 epochs and the valuation process of the model was done by Train Loss and Train Accuracy for each epoch of training and then Validation Loss and Validation Accuracy were calculated at the end of each epoch. After completion of training, the model was used to predict Test Audio files and assess the Precision, Recall, and F1-score of the model for each dataset. Along with accuracy score estimation, a confusion matrix for each emotion was sketched to check the performance of each model. 7 %. The training and testing accuracy and loss graph, confusion matrix, and classification report are shown in Fig. 10 - 12

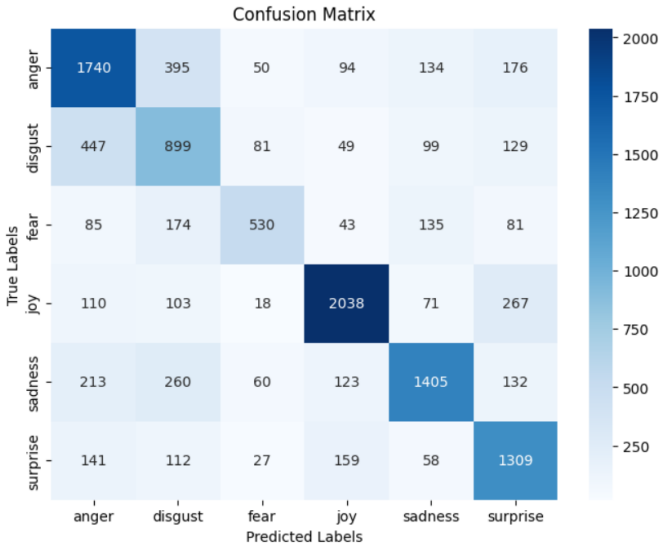

Fig. 8. Confusion Matrix for the test set of the RoBERTa model on GoEmotions dataset

	precision	recall	f1-score	support
anger	0.541330	0.801854	0.646326	2589.000000
disgust	0.570686	0.322183	0.411853	1704.000000
fear	0.638889	0.614504	0.626459	1048.000000
joy	0.839563	0.766782	0.801524	2607.000000
sadness	0.700452	0.636571	0.666985	2193.000000
surprise	0.692308	0.677741	0.684947	1806.000000
accuracy	0.660249	0.660249	0.660249	0.660249
macro avg	0.663871	0.636606	0.639682	11947.000000
weighted avg	0.671185	0.660249	0.654637	11947.000000

Fig. 9. Classification Report for the test set of the RoBERTa model on GoEmotions dataset

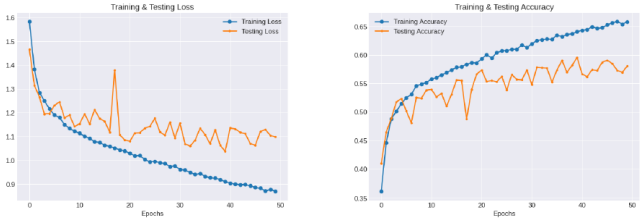

Fig. 10. Training and Testing accuracy and loss graphs for the test set of the CNN model

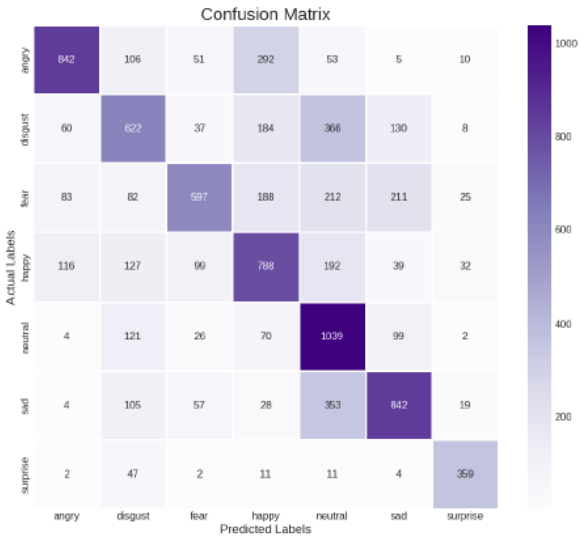

Fig. 11. Confusion Matrix for the test set of the CNN model

C. Facial Emotion Recognition Model

The model was tested and trained on the FER-2013 dataset. The CNN model for Facial Emotion Recognition was evaluated on 20% of the dataset. The overall accuracy of the model was 63.7% and the F1 score stayed at 63.3%. The confusion matrix and reports are shown in Fig 13.

D. Late Fusion Evaluation

The model was evaluated on MELD and IEMOCAP datasets. On MELD dataset the achieved accuracy was 33% and on IEMOCAP dataset it achieved 45% accuracy. The number of emotions detectable in MELD dataset is 7, while IEMOCAP was used to identify 5 exact emotions used for unimodal model classification tasks (Fig. **??** .

E. Web application testing

The web application was hosted in localhost of an Asus TUF Gaming FX505DT laptop with AMD Ryzen 5 3550H as CPU, GeForce GTX™ 1650 GPU, and 8 GB RAM. The average processing time for a 30-second HD video on this laptop was around 60-90 seconds when using only the CPU. The final size of the project was 5,33 GB and additional space is needed to store the uploaded videos. Despite having low performance on the benchmark dataset, with the custom dataset on real-world testing scenarios the results were reasonable. The output clips and statistics showed that the videos could mostly correctly identify the emotions experienced by the narrator. The test users also gave positive feedback on the user interface, as it provides clear information about the projects and provides all necessary instructions. In addition, intuitive design made it easy to understand and analyze the output results. As a result,

TABLE I
COMPARISON OF TESTING RESULTS FOR DIFFERENT DATASETS

Modality	**Dataset**	**Metrics**	**State-of-Art Result**	**Testing Result**
Image	FER 2013	Accuracy	76.12 %	64.7%
Text	MELD	Weighted F1-score	65.8%	62%
	GoEmotions	Weighted F1-score	55.8% (28 classes)	65.4% (7 classes)
Speech	Ravdess	Accuracy	84.75%	79.66%
	TESS	Accuracy	97.98%	97.18%
	CREMA-d	Accuracy	53.4%	50.75%
	All datasets together	Accuracy	–	82.42%

	precision	recall	f1-score	support
angry	0.76	0.62	0.68	1359
disgust	0.51	0.44	0.48	1407
fear	0.69	0.43	0.53	1398
happy	0.50	0.57	0.53	1393
neutral	0.47	0.76	0.58	1361
sad	0.63	0.60	0.62	1408
surprise	0.79	0.82	0.81	436
accuracy			0.58	8762
macro avg	0.62	0.61	0.60	8762
weighted avg	0.60	0.58	0.58	8762

Fig. 12. Classification Report for the test set of the CNN model

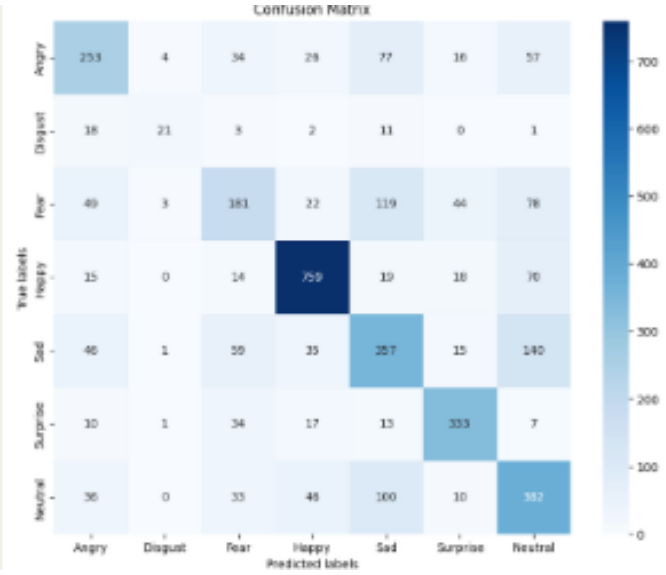

Fig. 13. Confusion matrix for the test set of FER for the CNN model

the functionalities provided by the app were highly evaluated by the testers.

VII. Conclusion and Possible Future Work

Integrating text, images, and audio data for emotion recognition yielded higher accuracy compared to unimodal approaches. However, significant challenges have arisen, primarily due to resource limitations, including limited computing power to process extensive data sets and a lack of comprehensive data sets covering all three methods. Existing datasets are primarily designed for sentiment analysis or do not have the necessary diversity for comprehensive testing, especially in video-based scenarios.

In the future, we will focus on improving fusion methodologies at both the decision-making and feature extraction levels. Improved fusion methods aim to optimize the integration of information across different modalities, thereby improving classification accuracy and reliability. Additionally, there is a concerted effort to incorporate physical data streams that complement the multimodal framework with contextual cues beyond traditional text, image, and audio data.

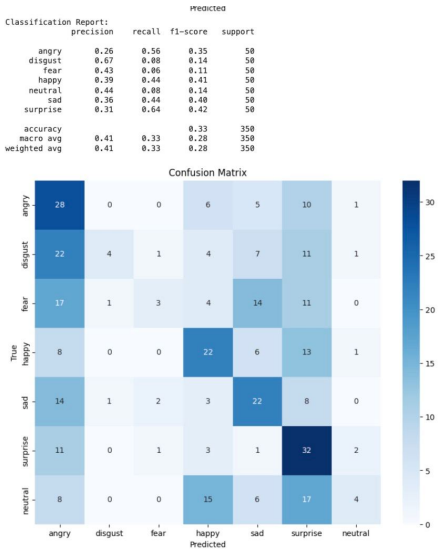

Fig. 14. Late fusion confusion matrix for

Beyond traditional applications such as interview analysis or sentiment scoring, we aim to leverage these advances in broader areas including medical diagnostics and personalized experiences. By seamlessly integrating physiological signals into the analysis pipeline, we aim to unlock new dimensions of precision and utility, driving innovation across multiple research areas and applications.

References

[1] R.W. Picard, Affective computing. MIT Press, 1997.

[2] A. Agrawal, R. Anil George, S. S. Ravi, S. Kamath S, and A. Kumar, "ARS$_{\text{NITK}}$ at MEDIQA 2019:Analysing Various Methods for Natural Language Inference, Recognising Question Entailment and Medical Question Answering System," Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-5059.

[3] X. Zhang, M.J. Wang, X-Da Guo, Multi-modal emotion recognition based on deep learning in speech, video and text, Proceedings of the IEEE 5th international conference on signal and image processing (ICSIP),2020, pp. 328-333

[4] M. I. Georgescu, R. T. Ionescu and M. Popescu, "Local Learning With Deep and Handcrafted Features for Facial Expression Recognition," in IEEE Access, vol. 7, pp. 64827-64836, 2019, doi: 10.1109/ACCESS.2019.2917266.

[5] A. Agrawal, N. Mittal, Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy, Vis. Comput. 36 (2) (2020) 405–412.

[6] D. Issa, M.F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomed. Signal Process. Control 59 (2020) 101894.

[7] E. Batbaatar, M. Li, K.H. Ryu, Semantic-emotion neural network for emotion recognition from text, IEEE Access 7 (2019) 111866–111878.

[8] A. F. Adoma, N. -M. Henry and W. Chen, "Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition," 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 2020, pp. 117-121, doi: 10.1109/ICCWAMTIP51612.2020.9317379

[9] D. Liu, Z. Wang, L. Wang, and L. Chen, "Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning," Frontiers in Neurorobotics, vol. 15, July 9, 2021, Article 697634. https://doi.org/10.3389/fnbot.2021.697634.

[10] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J.M. Montero, and F. Fernández-Martínez, "A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset," Appl. Sci., vol. 12, p. 327, 2022. https://doi.org/10.3390/app12010327.
[11] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative Multimodal Emotion Recognition Using Facial, Textual, and Speech Cues," 2019.
[12] M. Sajid, M. Afzal, and M. Shoaib, "Multimodal Emotion Recognition using Deep Convolution and Recurrent Network," in 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 128-133, doi: 10.1109/ICAI52203.2021.9445262.
[13] M. Li, X. Qiu, S. Peng, L. Tang, Q. Li, W. Yang, Y. Ma, "Multimodal Emotion Recognition Model Based on a Deep Neural Network with Multiobjective Optimization," Wireless Communications and Mobile Computing, vol. 2021, Article ID 6971100, pp. 1-10, 2021. https://doi.org/10.1155/2021/6971100.
[14] L. Cai, J. Dong and M. Wei, "Multi-Modal Emotion Recognition From Speech and Facial Expression Based on Deep Learning," 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, pp. 5726-5729, doi: 10.1109/CAC51589.2020.9327178.
[15] X. Zhang, M. -J. Wang and X. -D. Guo, "Multi-modal Emotion Recognition Based on Deep Learning in Speech, Video and Text," 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2020, pp. 328-333, doi: 10.1109/ICSIP49896.2020.9339464.
[16] L. D. Sharma and A. Bhattacharyya, "A Computerized Approach for Automatic Human Emotion Recognition Using Sliding Mode Singular Spectrum Analysis," in IEEE Sensors Journal, vol. 21, no. 23, pp. 26931-26940, 1 Dec.1, 2021, doi: 10.1109/JSEN.2021.3120787.
[17] J. Ma, H. Tang, W. Zheng, and B.-L. Lu, "Emotion Recognition using Multimodal Residual LSTM Network," in Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France, ACM, New York, NY, USA, pp. 8. https://doi.org/10.1145/3343031.3350871.
[18] K. Bayoudh, R. Knani, F. Hamdaoui, et al., "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," Vis Comput, vol. 38, pp. 2939–2970, 2022. https://doi.org/10.1007/s00371-021-02166-7.
[19] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," arXiv preprint arXiv:1503.01800, 2015. https://arxiv.org/abs/1503.01800.
[20] B. Duc, E. S. Bigün, J. Bigün, G. Maître, and S. Fischer, "Fusion of audio and video information for multi modal person authentication," Pattern Recognition Letters, vol. 18, no. 9, pp. 835–843, 1997.
[21] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, "Multimodal end-to-end sparse model for emotion recognition," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, Online, 6–11 June 2021, pp. 5305–5316.
[22] Y.R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," Multimed Tools Appl, vol. 80, pp. 2887–2905, 2021. https://doi.org/10.1007/s11042-020-08836-3.
[23] P. Kumar, S. Malik, B. Raman, and X. Li, "Hybrid Fusion Based Interpretable Multimodal Emotion Recognition with Limited Labelled Data," arXiv preprint arXiv:2208.11450, 2023. https://arxiv.org/abs/2208.11450.
[24] P. Kumar, S. Malik, and B. Raman, "Interpretable Multimodal Emotion Recognition using Hybrid Fusion of Speech and Image Data," Multimed Tools Appl, 2023. https://doi.org/10.1007/s11042-023-16443-1.
[25] N. Ding, S.W. Tian, and L. Yu, "A Multimodal Fusion Method for Sarcasm Detection Based on Late Fusion," Multimedia Tools and Applications, vol. 81, no. 6, pp. 8597-8616, 2022.
[26] Affective MELD Dataset. https://affective-meld.github.io/.
[27] State-of-the-Art Emotion Recognition in Conversation on MELD. https://paperswithcode.com/sota/emotion-recognition-in-conversation-on-meld.
[28] A. Agrawal, R. Anil George, S. S. Ravi, S. Kamath S, and A. Kumar, "ARSNITK at MEDIQA 2019:Analysing Various Methods for Natural Language Inference, Recognising Question Entailment and Medical Question Answering System," Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-5059.

APPENDIX

TABLE II
CLASSIFICATION REPORT FOR TEST DATA OF BASELINE MODEL

Class	Precision	Recall	F1	Support
Joy	0.66	0.11	0.19	402
Sadness	0.80	0.02	0.04	208
Fear	1.00	0.00	0.00	50
Anger	0.56	0.01	0.03	345
Neutral	0.51	0.97	0.67	1256
Disgust	1.00	0.00	0.00	68
Surprise	0.63	0.28	0.39	281
Acc			0.52	2610
M avg	0.74	0.20	0.19	2610
W avg	0.60	0.52	0.40	2610

TABLE III
CLASSIFICATION REPORT FOR FINE-TUNED BERT

Class	Precision	Recall	F1	Support
Joy	0.57	0.60	0.58	402
Sadness	0.41	0.27	0.33	208
Fear	0.22	0.16	0.19	50
Anger	0.49	0.43	0.46	345
Neutral	0.75	0.81	0.78	1256
Disgust	0.34	0.16	0.22	68
Surprise	0.53	0.62	0.57	281
Acc			0.63	2610
M avg	0.48	0.44	0.45	2610
W avg	0.62	0.63	0.62	2610

TABLE IV
CLASSIFICATION REPORT FOR FINE-TUNED BERT WITH RANDOM OVERSAMPLING

Class	Precision	Recall	F1	Support
Joy	0.58	0.58	0.58	402
Sadness	0.38	0.28	0.33	208
Fear	0.17	0.14	0.15	50
Anger	0.52	0.43	0.47	345
Neutral	0.74	0.82	0.78	1256
Disgust	0.41	0.13	0.20	68
Surprise	0.50	0.59	0.54	281
Acc			0.63	2610
M avg	0.47	0.42	0.44	2610
W avg	0.62	0.63	0.62	2610

TABLE V
DATASETS FOR DIFFERENT MODALITIES

Modality	Dataset	Description	Labels
Image	FER 2013	More than 30,000 48 by 48 grayscale images of people's faces	happy, sad, angry, fearful, surprise, neutral and disgust
Text	MELD	More than 13,000 utterances from the TV show Friends	happy, sad, angry, fearful, surprise, neutral, and disgust
	GoEmotions	58k comments extracted from Reddit, with human annotations to 27 emotion categories or Neutral	27 emotion categories and neutral, including happy, sad, angry, fearful, surprise, and disgust
Speech	RAVDESS	This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440.	RAVDESS includes calm, happy, sad, angry, fearful, surprise, and disgust expression
	CREMA-D	CREMA-D is a data set of 7,442 original clips from 91 actors	The sentences were presented using Anger, Disgust, Fear, Happy, Neutral, and Sad labels
	TESS	There are 2800 data points (audio files) in total	Recordings were made of anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral labels

TABLE VI
FRONT-END TOOLS AND LIBRARIES USED IN THE PROJECT.

Tools / Libraries	Version	Description
React	18.2.0	A JavaScript library for building dynamic user interfaces with a component-based architecture, virtual DOM, JSX syntax, and extensive ecosystem support.
Typescript	-	A superset of JavaScript that adds static typing to the language.
Bootstrap	5.3.2	Front-end framework for building responsive, visually appealing websites with predesigned UI components.
React-bootstrap	2.10.0	Bootstrap's UI components are implemented as React components for easy integration into React applications.
Axios	1.6.7	Promise-based HTTP client for making asynchronous HTTP requests in browsers and backend.
react-router-dom	6.22.1	Library for declarative routing in React applications, enabling navigation between views within single-page apps.
Chart.js	4.4.1	JavaScript library for creating customizable charts for data visualization.
React-chartjs-2	5.2.0	A React wrapper for Chart.js, allowing easy integration into React applications for data visualization.
React-icons	5.0.1	A library that offers a collection of popular icon packs as React components.
Redux	5.0.1	Predictable state container for managing application state in JavaScript apps.
React-redux	9.1.0	Official Redux binding for React, providing integration of Redux state management into React applications.

TABLE VII
MODEL INTEGRATION TOOLS AND LIBRARIES USED IN THE PROJECT.

Tools / Libraries	Version	Description
Tensorflow	2.16.1	An open-source machine learning framework developed by Google for building and training machine learning models.
Keras	3.1.1	A high-level neural networks API written in Python, designed for easy and fast experimentation with deep learning models.
NumPy	1.25.2	A powerful Python library for numerical computing.
Scipy	1.11.2	An open-source library for scientific computing and technical computing in Python.
Librosa	0.10.1	A Python library for analyzing and processing audio data.
SpeechRecognition	3.10.3	A Python library that allows the transcription of spoken words from audio files into text.
Transformers	4.40.0	A state-of-the-art natural language processing (NLP) library developed by Hugging Face, offering pre-trained models like BERT, GPT, and RoBERTa.
Dlib	19.24.2	A modern C++ toolkit containing machine learning algorithms and tools for creating complex software in C++ to solve real-world problems.
OpenCV-python	4.8.0.76	A Python library for computer vision and image processing tasks.
Imutils	0.5.4	A convenience library built upon OpenCV, offering a collection of utility functions for common image processing tasks in Python.

TABLE VIII
BACKEND TOOLS AND LIBRARIES USED IN THE PROJECT.

Tools / Libraries	Version	Description
Flask	3.0.2	A lightweight and flexible Python web framework that provides tools and libraries to build web applications quickly and efficiently.
MongoDB	-	A NoSQL database management system that stores data in flexible, JSON-like documents.
Flask_jwt_extended	4.6.0	An extension for Flask that provides JSON Web Token (JWT) support for authentication and authorization.
hashlib	-	A Python library providing secure hash and message digest algorithms.
pymongo	4.6.2	A Python driver for MongoDB, allowing developers to interact with MongoDB databases.
moviepy	1.0.3	A Python library for video editing at a programmatic level.

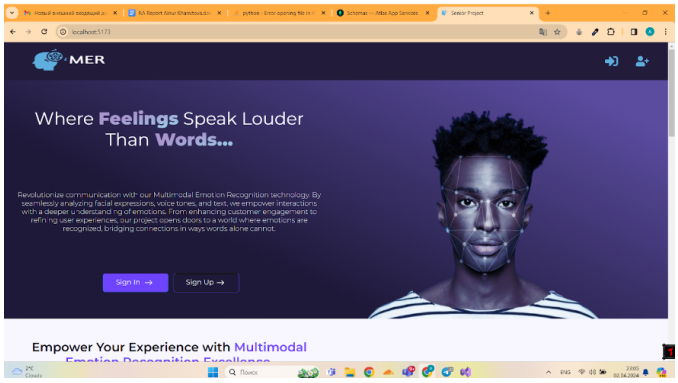

Fig. 15. The landing page of the web application

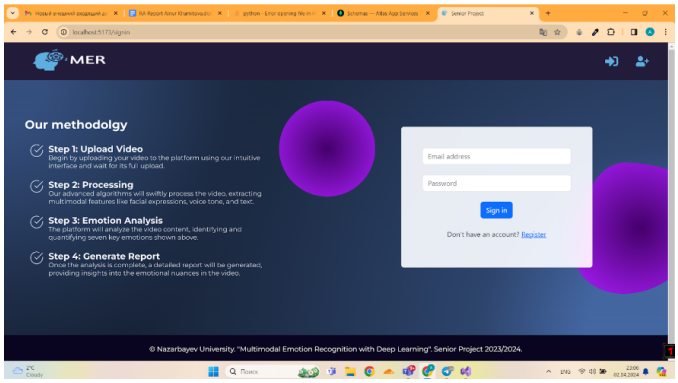

Fig. 16. The sign-in page

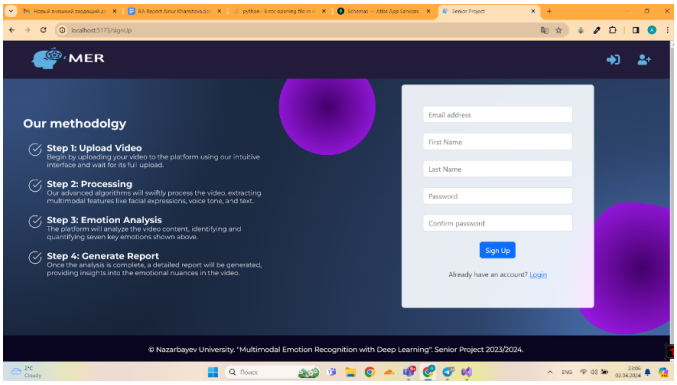

Fig. 17. The sign-up page

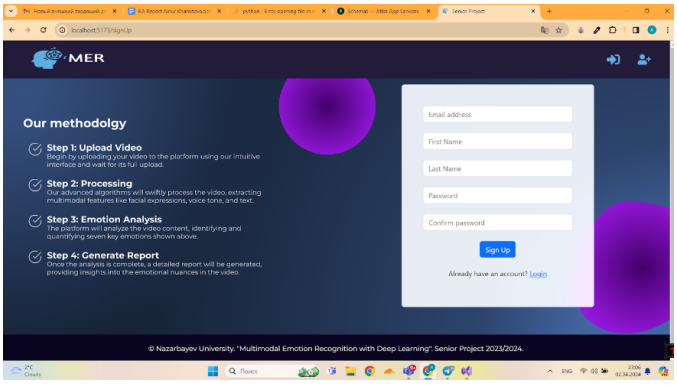

Fig. 18. The sign-up page

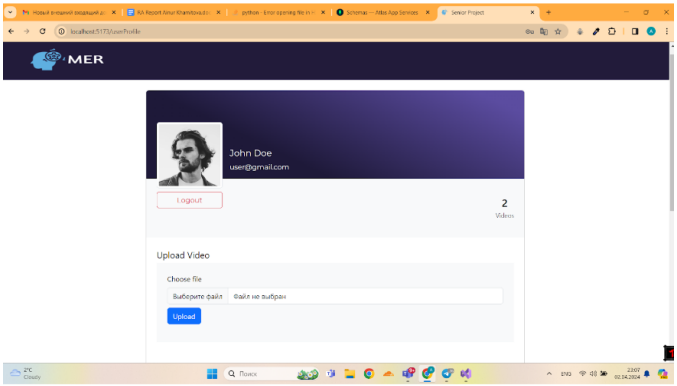

Fig. 19. The user cabinet

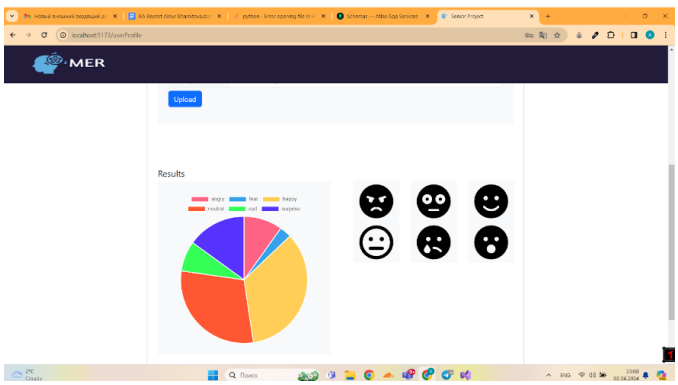

Fig. 20. The statistics and output emotions

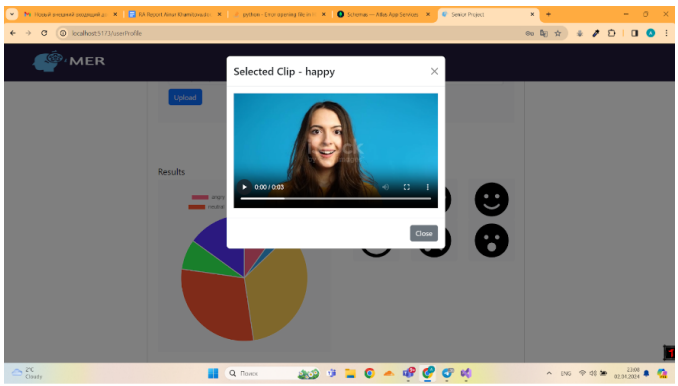

Fig. 21. Merged clip of present emotions with the same label

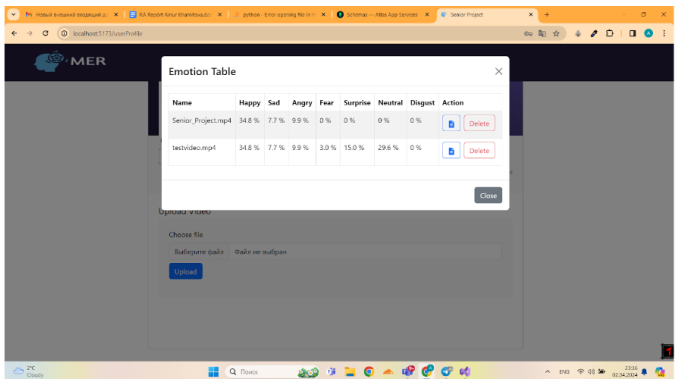

Fig. 22. History of results

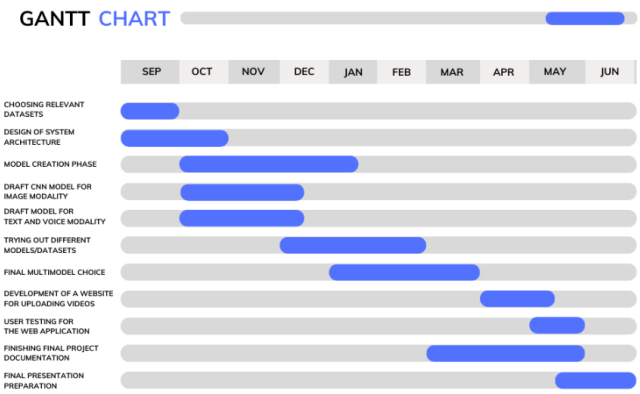

Fig. 23. Initial plan for the Fall and Spring Semesters

Neutral: What date is it today?

Fear: What's your biggest fear, and how do you cope with it?

Joy: What's one thing that never fails to make you smile?

Sadness: Can you recall a moment that deeply touched your heart?

Surprise: What was the most unexpected thing that ever happened to you?

Anger: What situation or event makes you feel the most frustrated or infuriated?

Disgust: Describe a time when you were utterly repulsed by something.

Fig. 24. Questions used for gathering the dataset