**Multi-Modal Emotion Detection System**

Group Link: <https://github.com/Oversoul73/DS-Capstone-Spring-2024>

Personal Branch Link: <https://github.com/Oversoul73/DS-Capstone-Spring-2024/tree/Mohmmad>

**INTRODUCTION**

In the period of human-computer interaction and full of feeling computing, understanding human feelings plays an urgent part in upgrading client involvement and engagement. To address this, our venture points to creating a modern multimodal feeling discovery framework leveraging facial expressions, voice sounds, and literary settings. By joining these different modalities, we look for a comprehensive arrangement capable of precisely recognizing and deciphering human feelings in real time. The system's backbone integrates the LLAMA (or LLAVA) model, a state-of-the-art deep learning architecture designed for multimodal learning tasks. Streamlit, a user-friendly web application framework, is used for the system's front end, providing an intuitive interface for users. MongoDB or Firebase are used for data storage and management, ensuring seamless integration with the emotion detection system. Flask is used for communication between the frontend and backend components, facilitating robust endpoints for data exchange. The project pushes the boundaries of emotion detection technology by leveraging synergies between facial, voice, and textual modalities. By integrating cutting-edge deep learning models, intuitive frontend interfaces, and robust backend infrastructure, the project aims to deliver a powerful multimodal emotion detection system capable of enriching human-computer interactions across various domains, including virtual assistants, social robotics, and affective computing applications.

**GOALS**

The objectives of our project can be summarized as follows:

1. **Create a Multimodal Emotion Detection System:**

Our essential objective is to form a vigorous framework that recognizes human feelings utilizing facial expressions, voice pitches, and printed inputs. By joining these three modalities, we aim to capture a comprehensive run of passionate prompts, empowering a more precise and nuanced feeling of acknowledgment.

1. **Utilize LLAMA, LLAVA and Wav2vec Model:**

We plan on using the LLAMA, LLAVA and Wav2vec models, that are profound learning structures optimized for multimodal learning errands. These models empower the successful combination of data from assorted modalities, improving the system's capacity to decipher human feelings precisely.

1. **Implement Streamlit Frontend:**

Our venture incorporates creating a user-friendly front-end interface utilizing Streamlit, a web application system. The front end will give natural visualization and interaction capabilities, permitting clients to input information and get real-time input on recognized feelings.

1. **Utilize MongoDB or Firebase for Data Storage:**

We proposed to utilize MongoDB or Firebase, adaptable NoSQL databases, for productive capacity and administration of multimodal information. These databases offer adaptability and execution, empowering consistent integration with our feeling discovery framework.

1. **Execute Flask APIs for Backend Communication:**

We are going to implement Flask, a lightweight web system, to construct strong APIs for communication between the frontend and backend components of our framework. Flask will encourage information trade and handling, guaranteeing seamless integration and smooth usefulness of the complete framework.

1. **Enhance Human-Computer Interaction:**

Ultimately, our project aims to improve human-computer interaction and effective computing applications by providing an effective device to detect and respond to human emotions in real life. We aim to improve customer encounters and engagement in different spaces by enabling precise and nuanced emotional localization across different modalities.

**SCOPE**

The project focuses on detecting human emotions using facial expressions, voice intonations, and textual inputs. It involves the integration of the LLAMA, LLAVA and Wav2vec models, advanced deep learning architectures optimized for multimodal learning tasks. Streamlit, a web application framework, is used for the front-end interface, but its complexity and sophistication may be limited by its capabilities. MongoDB or Firebase are used for efficient storage and management of multimodal data, but their scalability and data retrieval capabilities may be limited based on the dataset's size and complexity. Flask APIs for backend communication is implemented.

Limitations in emotion recognition accuracy may be inherent, as variability in human emotional expressions, cultural differences, and noise in input data could impact the system's ability to detect and interpret emotions accurately. Real-time processing constraints may also affect the system's processing speed and latency, as factors such as computational resources, network bandwidth, and input data complexity may influence the system's performance. Overall, the project aims to develop a multimodal emotion detection system that can effectively utilize various modalities.

**DATA SOURCES**

The multimodal emotion detection system will be trained using diverse and annotated datasets from Kaggle. Facial expression datasets will be used to identify different emotional states based on facial expressions like anger, sadness, etc. Voice intonation datasets will be used to capture audio recordings of individuals speaking with different emotional tones. Sentiment analysis datasets, emotion-labeled text corpora, and dialogue datasets containing emotional contexts will be explored to extract semantic features and contextual information to recognize the related emotions. Some links to datasets that we intend to use are as follows:  
<https://www.kaggle.com/datasets/sujaykapadnis/emotion-recognition-dataset>

<https://www.kaggle.com/datasets/uldisvalainis/audio-emotions>

<https://www.kaggle.com/datasets/zaber666/meld-dataset>

<https://www.kaggle.com/datasets/dileepathe/emotion-dataset>

<https://www.kaggle.com/datasets/robertknuth/emotion-dataset-aaai16>

<https://www.kaggle.com/datasets/omagarwal2411/nor-smart-speech>

Data augmentation techniques will be employed to increase the diversity and robustness of the datasets. Preprocessing will ensure consistency in format, resolution, and quality across different modalities.

Dataset integration and annotation will be done once the datasets are obtained. They will be integrated into a unified database structure compatible with MongoDB or Firebase, annotated with appropriate emotional labels and metadata to facilitate supervised learning and evaluation of the emotion detection system. This approach ensures the system's ability to recognize different emotional states based on facial expressions.

**METHEDOLOGIES**

Our research focuses on utilizing deep learning techniques to fuse information from various modalities, such as facial expressions, voice intonations, and textual inputs. We plan to explore multimodal learning architectures like LLAMA, LLAVA and Wav2vec models, which effectively combine features extracted from different modalities to capture complementary information.

Our multimodal emotion detection system will be trained in a supervised learning framework using annotated datasets with ground truth labels for different emotional states. We will employ appropriate loss functions and optimization algorithms to refine the model parameters and improve performance.

To evaluate the performance of our emotion detection system, we will use metrics like accuracy, precision, recall, and F1-score across different emotional categories. We will also conduct cross-validation and hold-out validation to ensure the generalization capability of our models across diverse datasets and real-world scenarios.

**Proposed Timeline**

Here's a proposed timeline with key milestones and estimated completion dates for the development of our multimodal emotion detection system up to April 14th:

**January 22 - February 5 (2 weeks):** Project Planning and Dataset Acquisition:

* Define project goals, scope, and requirements.
* Identify and capture facial expressions, tone of voice, and text emotion datasets.

**February 6 - February 26 (3 weeks):** Data Preprocessing and Integration:

* Preprocessing and cleaning of ingested datasets, including data augmentation and normalization.
* Consolidate datasets into a unified database structure compatible with MongoDB or Firebase.
* Annotate datasets with sentiment labels and metadata for supervised learning.

**February 27 - March 26 (4 weeks):** Model Implementation and Training:

* Investigate and select appropriate deep learning architectures for multimodal fusion, such as LLAMA and LLAVA models.
* Design neural network architectures for facial expression, speech, and text modalities for supervised learning.

**March 27 - April 9 (2 weeks):** Frontend Development and API Integration:

* Develop a front-end interface using Streamlit for user interaction and visualization.
* Implements Flask API for communication between frontend and backend components.
* Integrate the trained model into your backend infrastructure for real-time emotion recognition.

**April 10 - April 14 (1 week):** Testing, Validation, and Documentation:

* Document the entire project, including methodologies, implementation details, and user guides.
* Prepare presentations or demos showcasing the functionality and capabilities of the system.
* Finalize project deliverables and ensure readiness for presentation or deployment.

**My Contributions to the Project until now**

* I have researched datasets for the Audio modality that are listed in the data sources above.
* I have researched the Wav2vec model to process and train for the audio modality and helped in planning the direction of the project.
* Helped create the presentation slides for the Project and imbedded them to the README file in GitHub.