**EmotiSense: Multi-Modal Emotion Detection Framework**

**Summary Report**

The EmotiSense project aims to address the complexity of emotion recognition by building a multimodal model that combines data from three main sources: images, audio, and text. Traditional emotion recognition systems often struggle to capture the subtleties of human emotion because they rely on only one type of data. EmotiSense aims to overcome this limitation by leveraging the strengths of open-source Large Language Models (LLMs) such as BERT, LLaVa, and Wav2vec.

By integrating information from multiple modalities, EmotiSense can provide a more comprehensive and accurate understanding of human emotions. By analysing visual cues, tone of voice, and text content, the model attempts to capture the subtle nuances of emotional expression. This multimodal approach allows EmotiSense to identify emotions more effectively than systems based on only one type of data, making it a promising solution for applications that require robust emotion recognition capabilities. The main goal of this project is to push the boundaries of emotion recognition technology by leveraging synergies between different modalities. Through thorough experimentation and model refinement, EmotiSense aims to achieve superior performance in accurately recognizing and interpreting human emotions in various situations and domains.

The LLaVa model leverages feature extraction techniques, and CNN to extract relevant features from facial expression images and improve performance in real-time emotions. Achieve fast convergence and excellent performance, recognition task. Overall, the LLaVa model provides a robust framework for facial emotion recognition via image processing and significantly improves the accuracy of emotion recognition.

To implement Wav2Vec models for emotion recognition for speech analysis, you need to leverage pre-trained models such as Wav2Vec 2.0 and fine-tune them for emotion recognition tasks. This approach extracts features from audio signals, matches them to text-based language model, and uses a fully connected neural network classifier to accurately predict emotions. This process involves training a model on audio data, decoding the model output using a tokenizer, and optimizing the training process using a weighted categorical cross-entropy loss technique and ADAM optimizer. Overall, the implementation of the Wav2Vec model in speech analysis for emotion recognition represents a comprehensive approach to understanding emotion within audio signals.

The performance of the model with the initial training and validation is 14%. But after some initial (25 epochs, only the last 5 were shown in the technical document) fine-tuning the model it gave the accuracy at 17%. The training loss after 25 epochs is 11.8804, training accuracy is 16%. The validation loss is approximately 2 and validation accuracy 15%.

The workflow for using BERT for emotion recognition in text analytics includes pre-training the model on a large corpus of text data, fine-tuning it on emotion-labeled datasets, and understanding the context to effectively predict emotions.

We used ‘Emotions’ dataset which is a collection of English Twitter messages for text analysis. Each passage in this dataset comprises of a content fragment speaking to a Twitter message and a comparing name demonstrating the most feeling being passed on. Feelings are separated into six categories: pity (0), delight (1), adore (2), outrage (3), fear (4), and shock (5). When classifying feelings, this dataset gives a comprehensive establishment for exploring the nuanced enthusiastic scene of social media. The dataset is of 416808 records where 80% of the it is used for training and remaining for testing evaluation. We run the training epochs to 3, weight decay is set to 0.01.

In summary, we are in the process of fine-tuning for text and audio modality. We have changed our approach to text modality due to hardware and time limitation for now. We will start the process of finetuning for image modality within the following days.

EmotiSense could have a major impact on a variety of applications, including sentiment analysis, affective computing, human-computer interaction, and mental health monitoring. By providing deeper insights into human emotions, EmotiSense has the potential to enhance user experiences, improve communication systems, and drive advances in emotional intelligence technology.

In summary, EmotiSense is a promising effort in the field of multimodal emotion recognition, poised to address critical challenges and open new avenues for understanding and interpreting human emotions in diverse environments.