Multi-Modal Emotion Detection System

Integrating LLaMa, LLaVa,Wav2vec ...

Team Members:

Krishnasai Bharadwaj Atmakuri – bka2bg@umsystem.edu

Mohammadreza Akbari Lor – ma7fy@umsystem.edu

Mohitha Lakshmi Dayana – mdhkc@umsystem.edu

Hema Nagini Matta – hm2np@umsystem.edu

Problem Statement

The ability to accurately detect and interpret human emotions is a cornerstone in enhancing human-computer interaction. Traditional emotion detection systems have predominantly relied on single modalities, such as text analysis or facial recognition, leading to limitations in comprehensiveness and accuracy. The challenge arises from the complexity of human emotions, which are often conveyed through a combination of facial expressions, tone of voice, and verbal cues. Relying on a single modality can result in a significant loss of context and nuance, leading to misinterpretations. Furthermore, in a world that increasingly interacts digitally, the need for sophisticated emotion detection that mirrors human empathy and understanding has become paramount. This creates a significant opportunity to develop a more holistic, multimodal approach to emotion detection, leveraging advancements in large language models (LLMs) and machine learning.

Project Overview

This project aims to address these challenges by developing a multimodal emotion detection model that integrates data from three primary sources: image, audio, and text. By combining the strengths of open-source Large Language Models (LLMs) like LLaMa, LLaVa, and Wav2vec, this model seeks to provide a more comprehensive and accurate representation of human emotions. The integration of these modalities presents a unique opportunity to capture the subtleties and complexities of emotional expression in a way that single-modality systems cannot.

Project Overview Continued

Image Processing: Using LLaVa for the visual modality, the model will analyze visual cues from facial expressions and body language.

Audio Analysis: Wav2vec and similar tools will be employed to interpret vocal tonality, pitch, and other auditory signals that indicate emotional states.

Textual Interpretation: LLaMa will be used to understand the context and sentiment of spoken or written words, providing crucial insights into the emotional content.

The synergy of these modalities under a unified framework will enable the system to interpret emotions with a level of depth and precision akin to human interaction. This project is not just about advancing technology; it's about bridging the gap between digital interactions and human empathy, opening new frontiers in user experience, mental health, customer service, and beyond.

Potential Applications and Benefits

Healthcare: In mental health care, this technology could aid in monitoring patient's emotional states, providing valuable data for therapists and psychiatrists. It could also be used in telemedicine platforms to enhance patient-doctor communication.

Customer Service: In customer service, the ability to detect and respond to customer emotions can improve service quality. Automated systems and chatbots equipped with this technology could provide more empathetic and personalized responses.

Market Research: The technology can also play a significant role in market research, where understanding consumer emotional responses to products and advertisements can provide valuable insights.

LLMs (Large Language Models):

LLaMa: This model will be primarily used for textual data analysis.

Wav2vec: This model is crucial for audio data analysis. It will be used to process and understand speech patterns, tonality, and other auditory cues indicative of emotional states.

LLaVa: This model will be primarily used for visual data analysis and image processing.

A significant aspect of our data sourcing strategy involves the integration of these modalities. We will develop a framework to synchronize data from text, audio, and image sources, ensuring that they can be effectively combined for multimodal analysis.

Early datasets that are going to be used for fine-tuning the models

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

General Data Science Approach

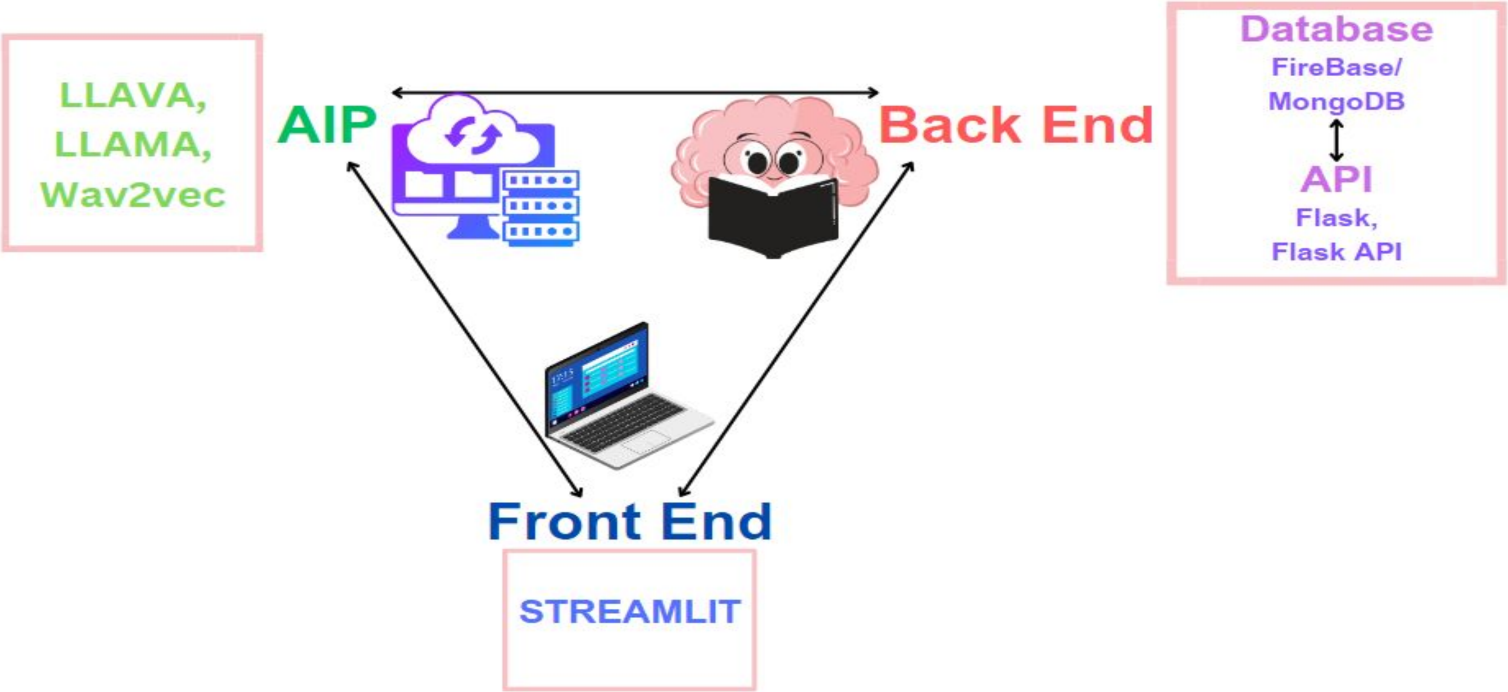
Data Collection and Preprocessing: Collecting data from various sources as previously outlined. Preprocessing involves cleaning, normalizing, and structuring the data for analysis. This includes noise reduction in audio data, image normalization, and text tokenization.

Feature Extraction: Identifying and extracting relevant features from each modality. For instance, extracting key facial landmarks from images, spectral features from audio, and semantic features from text.

Model Training and Validation: Applying machine learning algorithms to train the emotion detection model. This includes supervised learning techniques using labeled datasets. Validation involves assessing the model's performance using a separate dataset to ensure accuracy and reliability.

Fine-Tuning and Optimization: Fine-tuning the model parameters for optimal performance. This could involve hyperparameter tuning and testing different model architectures.

Architecture Diagram of Multi-Modal Emotion Detection System



Integrating LLMs

Integration of Textual Data with LLaMa: Utilizing LLaMa for advanced sentiment analysis and natural language understanding. Techniques like transfer learning can be employed, where pre-trained models are further fine-tuned with our specific datasets.

Audio Processing with Wav2vec: Implementing Wav2vec for extracting meaningful features from audio data. Focus on capturing emotional cues like tone, pitch, and speech rhythm.

Integration of Visual Data with LLaVa: Utilizing LLaVa for Image processing and emotion detection from faces.

Synchronizing LLM Outputs with Other Modalities: Developing a framework for integrating the outputs of LLMs with image and audio data analysis. Ensuring that the data from different modalities aligns correctly and complements each other in the final emotion prediction.

Challenges and Considerations in Multimodal Modeling

Data Synchronization and Fusion: Addressing the challenge of synchronizing data from different modalities, which may have varying formats, scales, and temporal resolutions.

Model Complexity and Computational Resources: Managing the increased complexity in multimodal models which require significant computational resources. Balancing model complexity with efficiency, especially for real-time applications.

Handling Data Imbalance and Bias: Addressing potential biases in the training data which can skew the model's performance. Implementing techniques to handle imbalanced datasets, such as resampling methods or cost-sensitive learning.

Interpretability and Explainability: Ensuring the model's decisions are interpretable and explainable, which is crucial for trust and ethical considerations. Implementing methods to visualize and explain the model's decision-making process, particularly how it integrates information from different modalities.

Expected Outcomes

Accurate Emotion Recognition: Enhanced ability to accurately recognize and interpret a wide range of human emotions using the combined data from image, audio, and text sources.

Contextual Understanding: Gaining deeper insights into the context behind emotional expressions, thanks to the multimodal approach which considers various aspects of human interaction.

Real-time Analysis: Developing a system capable of performing real-time emotion detection, which could revolutionize interactions in various fields.

Evaluation Criteria

Accuracy and Reliability: Metrics such as precision, recall, and F1 score in emotion classification tasks.

Real-Time Performance: Evaluating the latency and computational efficiency of the system in real-time applications.

Scalability and Robustness: Ability to maintain performance across diverse datasets and in different operational environments.

Conclusion

This project aims to develop a cutting-edge multimodal emotion detection system, leveraging the power of LLMs like LLaMa, LLaVa, and Wav2vec.

It stands at the intersection of technology, psychology, and user experience, aiming to bring a deeper understanding of human emotions to digital interactions.

The End