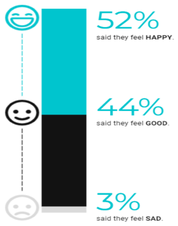
EmotiSense

A Multi-Modal Emotion Detection Framework

Mohammad Reza Akbari Lor, Krishnasai Bharadwaj Atmakuri, Mohitha Dayana, Hema Nagini Matta

Abstract

In the realm of human-computer interaction, accurately discerning and understanding human emotions plays a pivotal role. However, conventional emotion detection systems have predominantly relied on single modes of data analysis, such as text or facial recognition, resulting in limited accuracy and depth. The intricacy of human emotions, expressed through facial expressions, tone of voice, and language nuances, poses a significant challenge to these unimodal systems, often leading to misinterpretations and context loss. As digital interactions become increasingly prevalent, the necessity for advanced emotion detection systems that reflect human empathy and comprehension grows.



Multimodal emotion detection systems, driven by the need to capture the complexity of human emotions in digital interactions authentically. By integrating insights from facial expressions, vocal tones, and textual cues, our proposed approach aims to surpass the limitations of traditional unimodal systems. Leveraging advancements in LLMs and machine learning, our system seeks to enhance accuracy and contextual understanding, thereby facilitating more empathetic and nuanced human-computer interactions.

Introduction

Consider the challenge faced by mental health professionals in accurately assessing and monitoring the emotional well-being of individuals, particularly in remote or underserved communities. One major issue in this
context is the limited access to timely and comprehensive emotional assessments, leading to delays in intervention and support for those in need. Solution with EmotiSense, emerges as a transformative solution to this pressing problem. By leveraging its multi-modal emotion detection framework, EmotiSense can provide remote emotional assessments with unparalleled accuracy and efficiency.

Through the analysis of text, speech, and facial expressions, EmotiSense offers a holistic understanding of an individual's emotional state, even in situations where direct interaction with a mental health professional may not be feasible. Imagine a scenario where a person living in a rural area lacks access to regular mental health services. EmotiSense can serve as a virtual emotional companion, capable of detecting subtle changes in emotional patterns through everyday interactions such as text messages, phone calls, or video conferences.

Methodology

Data:

Here, these are our different modality datasets including the voice, text and image datasets which we used to train our models to detect the emotions

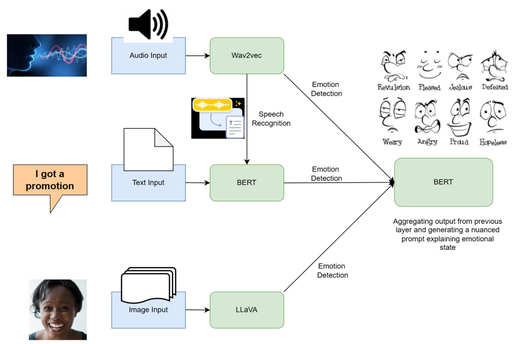
Text Dataset:
https://www.kaggle.com/datasets/parulpandey/emotion-datasetthis
Audio Dataset:
https://www.kaggle.com/datasets/uldisvalainis/audio-emotions
Image Dataset:
https://www.kaggle.com/datasets/robertknuth/emotion-dataset-aaai16

Data Cleaning and Transformation:

Voice Input:
- Resampling by reducing the frequency of the data
- Zero padding to have consistent audio length
- Converting audio into text using Wav2vec model

Text input:
- Feature extraction using the BERT model

Image input:
- Augmenting the data
- Feature extraction using the LLaVA model



As the interaction between all the platforms makes this system look great. We use Streamlit for frontend part. coming to API we use FastAPI. MongoDB is used for storing the data for training the models and take the input for determining the emotion of a person. Then we use LLMs for training the datasets.

LLMs (Large Language Models):

BERT: 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.

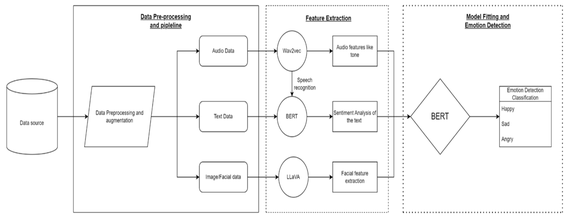
Integration of Textual Data with BERT: Utilizing BERT for advanced sentiment analysis and natural language understanding. Techniques like transfer learning can be employed, where pre-trained models are further fine-tuned with our specific datasets.

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.

Workflow:



Results

We have used BERT, LLAVa, Wav2Vec models respectively to predict the emotions in text, image, audio input.

The output given by LLAVa model:

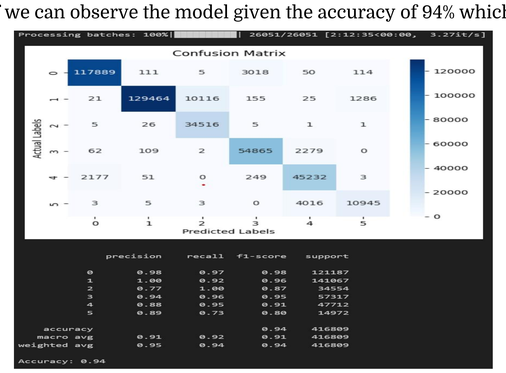
```
USER:  What is the emotion detected in this image?
ASSISTANT: The emotion detected in this image is happiness, as the woman is smiling brightly.

print(outputs3[0]["generated_text"])

USER:  What is the emotion detected in this image, give in only one word, the valid classification categories are: angry, disgusted, fearful, happy, neutral, happy, neutral, sad, surprised ?

ASSISTANT: Happy
```

The Confusion matrix for BERT model:
If we can observe the model given the accuracy of 94% which is



Future Work

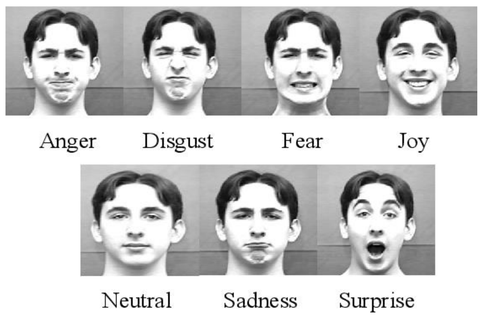
While EmotiSense represents a significant advancement in multi-modal emotion detection, there are several avenues for future exploration and enhancement:

1. Personalization and Adaptation: Explore methods for personalizing EmotiSense to individual users' unique emotional expressions and preferences, enabling more tailored and effective emotion detection and response.

2. Exploring Additional Modalities: Investigate the integration of additional modalities such as physiological signals (e.g., heart rate variability, skin conductance) and behavioral cues (e.g., gestures, body language) to capture a more comprehensive picture of human emotions.

Conclusion

In the pursuit of advancing human-computer interaction, our project endeavors to pioneer a groundbreaking multimodal emotion detection system that harnesses the capabilities of LLMs such as LLaMa, LLaVa, and Wav2vec. By synthesizing data from image, audio, and text modalities, we aim to transcend the limitations of traditional unimodal approaches and offer a more comprehensive understanding of human emotions in digital interactions. Throughout our endeavor, we have confronted various challenges, including data synchronization, model complexity, data imbalance, and interpretability, each of which we have addressed with innovative solutions to ensure the effectiveness and ethical integrity of our system.



References:

1. https://github.com/maelfabien/Multimodal-Emotion-Recognition
2. https://github.com/atulapra/Emotion-detection
3. https://www.nyu.edu/about/news-publications/news/2023/december/alexa--am-i-happy--how-ai-emotion-recognition-falls-short.html
4. https://www.mdpi.com/2504-2289/5/3/43
5. https://www.nature.com/articles/s42256-021-00417-9
6. https://www.comet.com/site/blog/ai-emotion-recognition-using-computer-vision/