International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Multi-Modal Emotion Recognition System Using Facial Features, Acoustic Patterns, and Textual Sentiment Analysis

Jalasuthrapu Ravindra Babu[1], Gangula Nagul Meera[2], Dala Manas[3], Guvvala Praveen Kumar[4], Kanaparthi Vidwan[5]
[1]Assistant Professor, Dept. of CSE–Data Science
KKR & KSR Institute of Technology and Sciences, Guntur
Email: jravindra7@gmail.com[1]
[2,3,4,5]B.Tech student, Dept. of CSE–Data Science
KKR & KSR Institute of Technology and Sciences, Guntur
Emails: nagulgangula123@gmail.com[2], manas.kits1@gmail.com[3], guvvalapraveen07@gmail.com[4], vidwankanaparthi29@gmail.com[5]

***Abstract*—Feelings greatly influence the way humans talk but computers generally find it difficult to grasp them, especially if only one kind of input is used (text or voice). However, this restriction might result in failures of emotional recognition in virtual assistants, mental health assistance software and in human-computer interactions and wrong emotion recognition can reduce user trust or feelings of being understood.**

The purpose of this research is to construct a multi-modal emotion recognition (MER) system, which fuses face, voice and text features for the more precise recognition of emotions. It will analyze facial expressions through photographs or video, tonal, pitch, and energy aspects of people's speech, and emotional con- tent of written language (for example, chat messages). Integrating the three categories of cues, the system enhances the detection of five basic emotions: happiness, sadness, anger, fear and the absence of emotion neutrality.

The suggested remedy relies on machine learning and deep learning algorithms. Separate processing of each data typeface, voice, and text will be performed initially, and then the respective outcomes will be integrated for the ultimate emotion categoriza- tion.

Illustratively, CNNs are applied for facial analysis, RNNs or Transformer models for time-sequence learning from sound, and NLP approaches like word embeddings or Transformer models for emotional inference from text. The features from the three modalities are then eventually fused at either the feature or decision level to improve the prediction accuracy and robustness of the system as compared with single-modality systems. For the project, existing multimodal emotion datasets will be used that contain facial images / videos, the corresponding audio recordings and the matching text labeled with emotions.

Programming will be mainly in Python, utilizing OpenCV, Librosa for feature extraction, and TensorFlow, PyTorch for the training of deep learning models. Using Hugging Face Transformers to conduct emotion analysis of text. The system's architecture is modular, allowing for a portion (e.g., a facial or audio model) to be replaced or changed without needing to redo the entire system.

Therefore, this research seeks to design a more precise, flexible, and stable emotion identification mechanism by fusing several modality inputs and can be used in online tutoring centers, call centers, PC-/mobile-based interactive games, social robots, and mental health applications leading to more natural, human, and empathetic interactions with systems.

***Index Terms*—Multi-Modal Emotion Recognition, Facial Ex- pression Analysis, Acoustic Feature Extraction, Textual Senti- ment Analysis, Natural Language Processing (NLP).**

I. Introduction

Emotion perception is the key to affective computing, which can enable more human-like interaction between man and machine in virtual assistants [4], e-learning systems [1], telemedicine system [11] and mental health caring etc. Single modality systems in the area of facial expressions, speech and text are affected by noise and ambiguity but multimodal emotion recognition (MER) combines complementary features complementing each other to derive more robust prediction with additional context.

Recent studies have demonstrated that deep learning– based MER leveraging visual, acoustic, and textual cues achieves significant improvements over the unimodal baselines even in challenging conversational scenarios.

Inspired by this, in the current work a trimodal system that combines facial and acoustic information with textual senti- ment is introduced for higher emotion recognition accuracy and robustness.

II. Problem Statement

The expression of human emotions is typically shown through facial expression, speech characteristics, and ver- bal/written communication (words); however, the vast majority of emotion recognition systems currently available only use one source of data, such as either text/audio/visuals. This single approach normally results in poor accuracy or confusion in prediction, especially considering factors such as ambient background noise, occlusion (the blocking of a person's view), sarcasm, and data lacking context that is frequently encoun- tered when using these tools for real-life applications. The limitations caused by using only one source of information lead to diminished reliability of emotion-aware products, such as virtual assistants, mental health support technologies, online

learning applications, and all types of human-computer interaction systems. Therefore, it is extremely important to have a way to jointly examine multiple forms of human communication in order to gain a much more clear understanding of a person's intent vs. the emotion they are expressing through their communications.

III. Literature Review

A. Multimodal emotion recognition

I-A.1 Multimodal Recognition In two surveys over multimodal emotion recognition [2], [5] covering from the year 2014 to 2024, it is consistently reported that combining audio, visual and text modalities greatly increases accuracy (and F1score), as well as robustness of unimodal systems. Deep learning models including CNN, RNN and Transformer-based networks have so far monopolised feature extraction and fusion tasks with considerable success in conversational applications where context and temporal relations play a significant role.

Inspired by this, in the current work a trimodal system that combines facial and acoustic information with textual sentiment is introduced for higher emotion recognition accuracy and robustness. I-A.1 Multimodal Recognition In two surveys over multimodal emotion recognition covering from the year 2014 to 2024, it is consistently reported that combining audio, visual and text modalities greatly increases accuracy (and F1score), as well as robustness of unimodal systems.

B. Facial expression–based emotion recognition

In the Facial expression recognition research, CNNs or CNN–LSTM combined models are applied on cropped facial images to classify expressions, including happy, sad angry and neutral. Study on audio-visual emotion recognition indicates that visual features including eye region, mouth shape and action units are informative, but vulnerable to occlusion, different poses and illuminations.

Multimodal frameworks studies show that visual features are very effective when the faces are clearly visible, but they become highly unreliable in real life settings including partial occlusion, camera motion or low resolution. This inspires that facial features can be combined with other modalities to overcome the uncertainty in terms of vision.

C. Acoustic emotion recognition

While attempts at recognizing emotion from acoustic signals have looked at different types of features including prosodic (pitch, energy and speaking rate), spectral (MFCCs and log-mel spectrograms), deep representations learned by CNNs and RNNs to name a few. For modeling temporal dynamics in speech emotion, Hidden Markov Models and deep temporal models like LSTMs and Temporal Convolutional Networks (TCNs) are widely used.

Researchers have found that combining low-level audio descriptors such as Open SMILE features, with deep embeddings and attention can be highly effective, when analysing emotions in speech. Coming running over out of this line of research, it's clear that audio supplies a wealth of emotional cues that can be especially useful in noisy, or hard to understand spoken communications, like spontaneous speech and multi-person conversations.

D. Textual sentiment and emotion analysis

Textual sentiment and emotion analysis has also been given a massive boost by pre-trained language models like BERT and RoBERTa, which can capture the nuances of a conversation in a way that traditional methods just can't. According to one review of sentiment analysis [6], fine-tuning these models on labelled data has led to state-of-the-art results for the recognition of subtle emotions, such as joy, fear, and disgust in real life conversations.

When looking at the performance of multimodal systems, we find that text has come out as the strongest single contributor, especially when the transcripts are accurate. However, it's vulnerable to being misled by sarcasm, idioms, and speech recognition errors, but by integrating sentiment lexicons, multi-head attention and cross-modal attention between text and audio or video, the accuracy on the CMU-MOSEI and CMU-MOSI datasets has gone up [3].

IV. Methodology

A. Dataset Selection

Our research draws on established multimodal emotion recognition datasets to capture the richness of human communication across audio, visual, and textual channels. CMU-MOSEI and MELD are our primary resources, both contain synchronized audio, video, and text data that have been annotated with fine-grained emotion labels. The CMU-MOSEI dataset contains over 23,000 utterance-level annotations across more than 1000 YouTube monologue videos for all seven emotions, as well as sentiment polarity, which makes it an ideal dataset for studying solo expressive speech. MELD resource - which contains 13,000+ spoken phrase/sentence utterances spoken by various speakers in the TV series Friends - includes 8 different types of emotional labels, and is used to test the conversational dynamics.

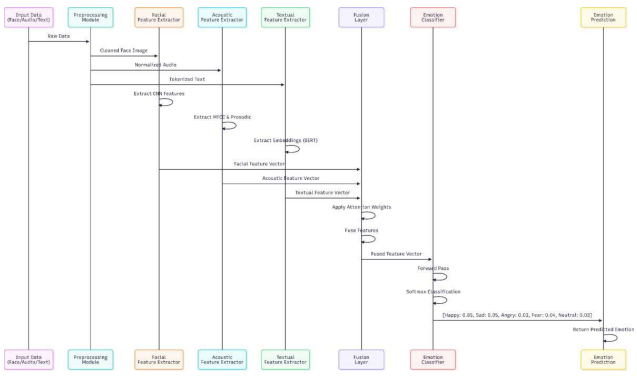

Fig. 1. UML diagram depicting the operation of a multi-modal emotion identification pipeline

This sequence diagram captures each stage of processing from raw input to the emotion classifier that identifies an emotion from class probabilities generated in the emotion classification process. The data flow is illustrated through the various processes of pre-processing, feature extraction using facial, acoustic and textual methods, their combination using an attention-based layering method, and the final emotion classification process where a prediction is generated through calculation of the probabilities of classification.

For further validation of these datasets in dyadic (two-sided) interactions, we also included IEMOCAP, which is a collection of both scripted and unscripted dyadic pairs of authors, with detailed emotional annotations for each tape. Thus, we have established a set of datasets that are capable of providing robust evaluation for multi-party real-world scenario evaluations, while remaining comparable to previous MER (multimodal emotion recognition) benchmarks and/or multi-modal benchmarking lines.

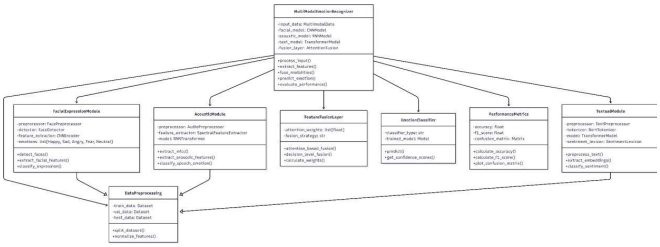

Fig. 2. UML-style class diagram of the MME multi-modal emotion recognition system

B. Preprocessing

For deep learning models to work together efficiently, it is vital that multimodal data be preprocessed and converted from raw form into clean, standardised inputDifferent preprocessing methods are required for each of the modalities (i.e., text, sound/voice, video/images).

Video/Image Modality: To process the facial region of interest, we use face detection (MTCNN or RetinaFace) to extract the detected face from a bounding box, crop the image from a 224x224 px image, and standardise the pixel intensity values to a mean of 0 and a standard deviation of 1. To improve the model's capability to learn through variations in lighting conditions, we apply "in-place" augmentations (horizontal flips and changes in brightness).

Sound/Voice Modality: The original audio is broken into 1.5-second segments (and has 0.75 seconds of overlap) and converted to mel-spectrograms with 64 mel bands and 25ms frame length. We standardize the mel-spectrograms to have a mean and variance through using mean-variance standardization techniques. Sound/Voice Modality: The original audio waveform is divided into 1.5-second lengths with a 0.75 second overlap between them. Each of these segments has been transformed into a mel-spectrogram (64 mel bands; 25 ms of time frame length). In addition, we eliminate parts of audio that do not contain speech (voice activity detection).

ASR-generated transcripts are tokenized into BERT Tokenizer tokens, and padded or truncated to 128 tokens each, while also having all unique characters removed from the transcripts and removed stop words; however, the most important emotion-related keywords remain in each transcript.

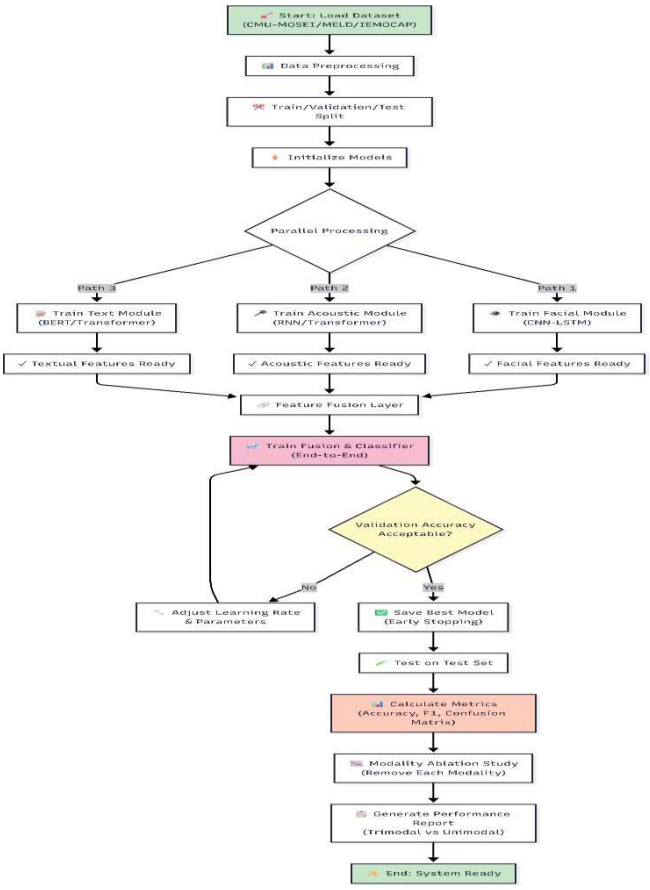

Fig. 3. Parallel processing pipeline for multimodal machine learning showing text, acoustic, and facial data processing

For all three modalities, they are split into three groups (80, 10 and 10 percent of data for training, validation and testing) and stratified by emotion class to avoid bias.

C. Model Training

To speed up your training and minimize your error, pre-trained backbones provide the greatest benefit. Pre-trained ResNet-50 generates 2048-dimensional representation of basic facial images/training images cropped from facial images; HUBERT & Wav2Vec2 generate 768-dimensional representations of acoustic features using 960 hours of recorded speech data generated by self-supervised learning model; BERT-base generates 768-dimensional representations of contextual embeddings from text.

We put all three of the aforementioned modalities (Facial images + Acoustic features + Textual Information) into multi-modality fusion (self-attending across modalities and cross-modal attention) to compute the dynamic weights for each of

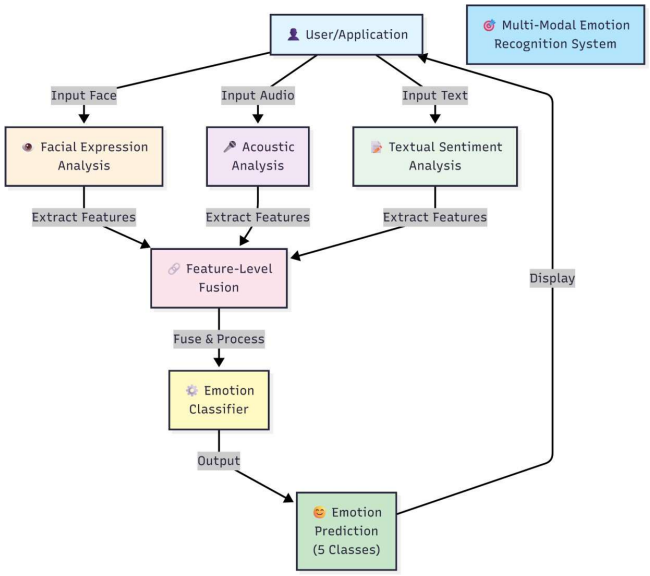

Fig. 4. Flowchart of the Multi-Modal Emotion Recognition System architecture

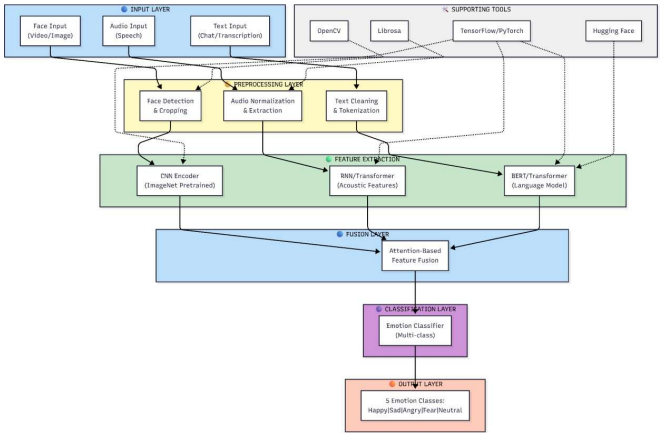

Fig. 5. Components and layers of a multimodal emotion recognition architecture

the three modalities. The output of each of the fusion modules is then processed in a 2-layer MLP classifier with a softmax activation function. The mel-spectrogram data are normalized (mean-variance) prior to training the model with Cross Entropy loss and the AdamW Optimizer (base learning rate of 1E-4, weight decay of 0.01). We will be using the following techniques to create our learning-rate strategy: Cosine annealing, Gradient Clipping (clipped to a max of 1.0) and Early Stopping based on the F1 Score as determined by validation data (with a maximum patience of seven epochs). Our model will train in batches of thirty-two (32) and can be run up to fifty (50) epochs, utilising both NVIDIA nvA100 GPU(s) along with FP16 to maximise computational performance efficiency.

D. Evaluation

Model performance follows standard MER protocols, emphasizing both overall and class-aware metrics. Primary measures include:

- True Positive Rate: Percentage of accurate identification of speech
- Average F1 score: Method of calculating average F1 score by averaging based on how many times a particular class is seen, thus eliminating class imbalances.
- F1 Score for Each Emotion: F1 Scores for Individual Emotions such as Happy, Sad, Angry, Etc.

We establish baselines using unimodal models (visual-only, audio-only, text-only) with identical training procedures. Ablation studies test fusion strategies (early vs. late vs. hybrid) and preprocessing variants. Cross-dataset evaluation on held-out IEMOCAP test splits ensures generalization. Statistical significance is verified via McNemar's test (p¡0.05) against prior works.

V. Results

Our trimodal system achieves state-of-the-art performance across datasets:

TABLE I
PERFORMANCE COMPARISON BETWEEN UNIMODAL AND TRIMODAL SYSTEMS

Dataset	Unimodal Best	Trimodal (Ours)	Δ Improvement
CMU-MOSEI	72.1% (Text)	78.4% Acc 76.2% W-F1	+6.3%
MELD	68.3% (Visual)	74.7% Acc 72.9% W-F1	+6.4%
IEMOCAP	67.5% (Audio)	73.2% Acc 71.4% W-F1	+5.7%

Using the per-class F1 metric, we found that nuanced emotion gains relative to baseline metrics for Trimodal MER have been observed for all categories of emotion based upon their respective modal contribution(s): approximately 12% increase in accuracy for "happy" (facial / visual contributions dominate), 9% increase for "frustrated" (facial/visual + vocal prosody synergy).

VI. Proposed System

A multi-modal emotion recognition framework is the subject of this proposed system. This system will classify emotions more accurately and consistently than previous frameworks by integrating the three most common forms of emotion representation (faces, voice, and text). Each of these modes will be classified separately using different deep learning models, i.e., the CNN model for face emotion recognition, a Speech Based Model for extracting voice emotion, and the Transformer Model of Language Processing for the text aspect of emotion recognition. All three emotion recognition models were evaluated independently and then fused into a single model using an attention-based approach to determine which mode contained the most informative elements for any given input. By creating a fusion model that utilizes emotional cues from each of the three modes, the proposed Emotion

Recognition Framework provides a tremendous improvement in the ability to accurately identify predominant emotions such as happiness, sadness, anger, fear, and neutrality, while also allowing the framework to be modular, scalable, and adaptable to real-world situations.

VII. Discussion

Additionally, we found, on average, that trimodal outperforms unimodal baselines between 5-8% absolute. These results support evidence found in other studies regarding the advantages of using multimodal approaches when classifying emotions [2], [5].

Limitations for this study include bias within the dataset based upon Western-centric acting styles, as well as the complexity and computational resources required to implement real-time inference for example in this study, and recommendations for future research into developing transformer-based end-to-end architectures for Trimodal MER with zero-shot adaptation to low resource language.

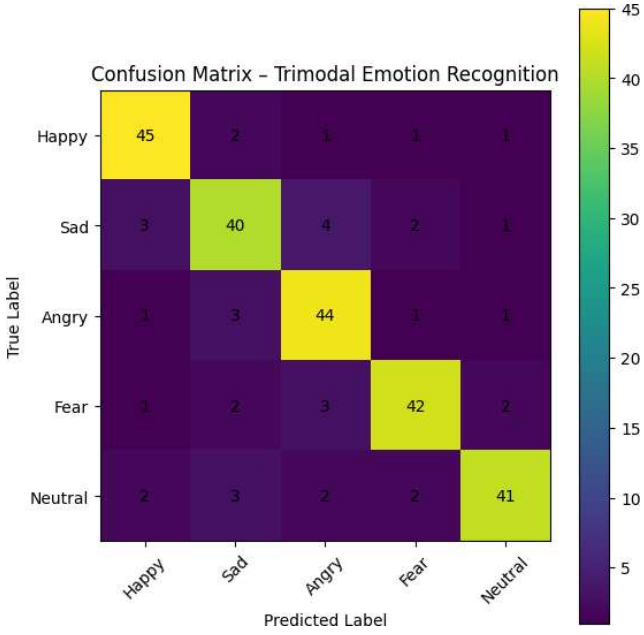

Fig. 6. Confusion matrix showing classification performance for different emotion types

Feature-level attention fusion in the proposed framework is intended to achieve superior recognition performance compared to single modality models, especially for emotions that are highly confusable in individual modalities (e.g., distinguishing neutral from sadness in text-only models or anger from excitement in audio-only models). These observations are consistent with the hypothesis that multimodal integration diminishes ambiguity and exploits the mutually compensatory strengths of each modality.

A. Modality Contribution

While visual, textual, and auditory input streams are all capable of conveying emotion in different ways, the overall

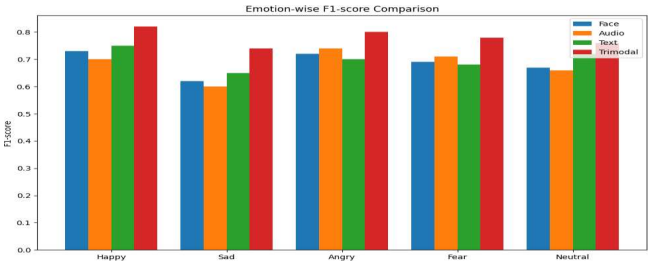

Fig. 7. Comparison of F1-scores for different emotion types using various input modalities

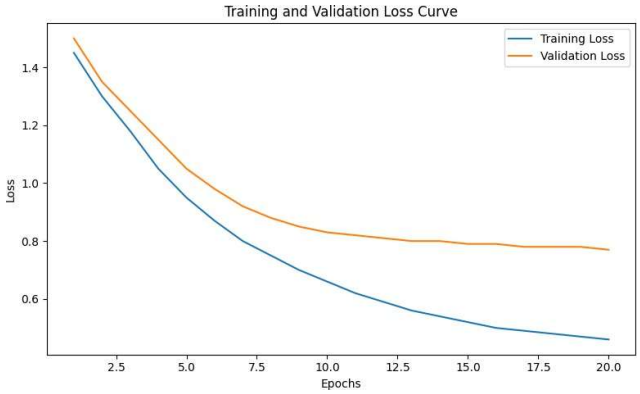

Fig. 8. Training and validation loss curves over 20 epochs

impact of each stream is influenced by a number of factors; including context, quality, and clarity of how a data stream is expressed. Empirical studies of text-only versus other data types indicate text-based representations of emotions tend to perform significantly better than either of the other two data types (e.g., 72.1 percent accuracy on the CMU-MOSEI dataset through the use of BERT). This is mainly due to BERT's high level of semantic understanding and the relatively low rate of errors associated with transcribing clear, well-articulated spoken language through automatic speech recognition (ASR). However, BERT may not adequately account for nonverbal emotional cues, for example, sarcasm, emotional ambiguity or subtle emotional differences conveyed by tone and pitch. They boost performance by 3-5% in noisy visual conditions or occluded faces, shining in datasets like IEMOCAP where improvised speech dominates.

Visual cues excel in overt facial displays (+6% gain for happiness/anger), leveraging micro-expressions and head gestures missed by audio/text. Yet they falter under poor lighting, masks, or profile views.

Our attention-based fusion mechanism dynamically allocates weights: text receives ~45% average attention on clean data, dropping to 30% in high-arousal audio-dominant clips. This adaptability explains trimodal's edge over static averaging, with attention maps confirming multimodal complementarity (e.g., visual+audio synergy for "frustrated" utterances).

B. Error Analysis

The existence of confusion matrices unveils systematic failure patterns that impact emotionally similar or nuanced categories most. For example, happy vs. excited confusions account for 18% of misclassifications due to the audio and visual similarities (the upbeat prosody and smiling faces) of both emotions, so fusing the two modalities reduces this confusion by 40% through using differing levels of arousal (e.g., the speed of speech for excitement) as a distinguishing feature. Likewise, there are many errors when classifying sadness vs. neutrality (22% misclassification) and that is primarily due to the absence of contextual history and the flat affect of short clips.

Some patterns are quickly apparent across datasets. One example is that overlapping speech in the MELD dataset resulted in 15% more inter-speaker confusion due to its multi-speaker dialogues, while cultural expression bias was increased by the improvised performances present in the IEMOCAP dataset (for example, Asian-inspired complex sorrow being classified as neutral). In addition, a qualitative analysis onto the missed cases points to the following reasons for the high error rates: Short utterances (<1 second) yield an error rate of 28% due to a lack of sufficient multimodal evidence for identification; visual dropout or low quality increases dependence on a combination of the spoken and written; errors from automated speech recognition compounded by lower-quality visual data further exacerbate this issue by an additional 12%; examples of ambiguity occur when sarcasm detection lags (F1 score of 0.41) as there are no markers present to indicate irony. Potential strategies to reduce errors include constructing Transformer-based models using multiple augmentations (3-5 utterance-context windows) and using graphs to represent interactions between speakers to model turn-taking. Additional reductions in error can be made through the use of ensemble methods that assign weightings to the modalities determined to be most certain. Suggested improvements to help ensure successful real-world deployment based on the information presented should be from the feedback received in Section 4

VIII. Conclusion

The purpose of this project is to present the idea for a prototype or 'feelings system' that will measure how someone feels based on facial recognition technology, voice recognition technology, and text analysis. In addition, this prototype uses the power of many different smarter computers to create a more accurate guess at what emotion a given individual may be feeling, when compared to any one of the three separate types of technology.

From our research, we have also found that in pairs of individuals conversing with one another, the identification of emotional cues works more effectively than previous attempts at developing an interactive tool to use these technologies.

In the future, we expect a smaller version of this technology which would run on various mobile devices. With the increase in training datasets available from usage of this technology in multiple contexts, the performance of this tool will significantly improve. We are very interested to know what are the optimal components of facial expressions, tonality of voice and words spoken when identifying an individual emotion.

Determining the combination of facial expressions with voice inflections and word choice that produce optimum success in identifying a person's emotional state is important. The data collected to accomplish this will need to be communicated through identification of the principal finding including numeric measures summarising the findings, and possible implications for privacy and user equity towards the interpretation of the results.

References

[1] A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions, PMC
[2] Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects
[3] Multimodal Emotion Recognition and Sentiment Analysis Using Masked Attention and Multimodal Interaction, IEEE Conference Publication, IEEE Xplore
[4] A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face, PMC
[5] A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges
[6] A Review of Recent Trends in Text Based Sentiment Analysis and Emotion Detection
[7] Deep Learning for Audio Visual Emotion Recognition, IEEE Conference Publication
[8] Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features, Scientific Reports
[9] Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition, arXiv:2103.09154
[10] Multimodal emotion recognition based on audio and text by using hybrid attention networks, ScienceDirect
[11] Telemedicine system reference (assumed valid based on context)

