Vibe Check: Multi-Modal Emotion Recognition at the Edge

Haley Lind, Joshua Moorehead, Regan Willis
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29208

***Abstract*—Our system combines a facial expression recognition network with a sentiment analysis network, fused at the last layer of the models. The output emotion is determined as either positive or negative.**

***Index Terms*—Facial Expression Recognition, Sentiment Analysis, Data Fusion, Edge Computing**

I. Introduction

As machines interface more frequently and intimately with people, their ability to understand and respond to human emotions in real-time becomes integral to their usefulness across a wide variety of applications, including health care, sales and marketing, and politics. In order to have a full understanding of human emotions, machine learning models must be able to see the complete picture, which includes verbal and non-verbal cues given by a person that provide insight into their emotional state. In order to realize this with machine learning, a multimodal approach may be taken to process the sentiment of a person's speech alongside the recognition of a facial expression at the time of speech.

We must consider the speed of the approach, as in certain applications emotions should be addressed sooner rather than later. Additionally, speech and video data may be private and therefore necessary to process locally. For these reasons, we deploy our system on an edge device and include processing speed as a metric in our evaluations along with accuracy.

Our implementation comprises two main components: a facial expression recognition model based on the PAtt-Lite architecture achieving 62.15% accuracy on the FER2013 dataset, and a BERT-based sentiment analysis model achieving 93.01% accuracy on our test dataset. These components are fused at the decision level to produce a final positive/negative emotion classification. The complete system was successfully deployed on a Raspberry Pi 5, with the FER model achieving an inference time of 1.4 seconds and the sentiment analysis model achieving approximately 313 milliseconds, making it suitable for real-time applications.

II. Literature Review

We investigate recent related works in the fields of facial expression recognition, sentiment analysis, and data fusion.

Facial Expression Recognition Facial expression recognition (FER) is a developing research area in computer vision that aims to categorize human emotions from visual facial features [1]. FER has extensive applications in healthcare, driver monitoring, security, and education. The conventional approach to FER followed a three-stage pipeline of face detection, handcrafted feature extraction, and classification using traditional machine learning methods. However, with the advancements of deep learning, these steps have been unified into end-to-end models with convolutional neural networks (CNNs) and more recently vision transformers (ViTs) to provide better performance and generalization. Even so, the implementation of FER faces difficulties when processing real-world data. Training images are typically burdened with occlusion, lighting changes, pose changes, and class imbalance. In addition, annotation inconsistency, which is when facial expressions are annotated ambiguously or incorrectly, can further weaken model performance. A growing research trend is engaged in putting FER models on edge devices, which poses tradeoffs of maintaining high accuracy and real-time inference while under computational constraints [1].

Real-time facial expression detection on low-resource edge devices demands the development of efficient yet accurate models [2]. Light-FER is a system tailored to address this requirement, with a suitable light adaptation of the Xception network as its foundation architecture [2]. In order to deploy the model efficiently on devices like the Jetson Nano, the authors employ pruning and quantization techniques, which reduce the number of parameters and the model size significantly [2]. The final model achieves 70.2% accuracy on the FER2013 dataset and requires just 4.3 MB of storage space [2]. This speed-accuracy-small model trade-off reveals that model compression achieves high performance under the edge computing constraint. The work reflects the growing concern in FER research on making deployment feasible versus recognition performance, particularly for low-power applications [2].

To facilitate efficient facial expression recognition at the edge, we selected the PAtt-Lite model as the baseline for our system. PAtt-Lite is a patch attention network that introduces a lightweight yet expressive architecture that is particularly designed for FER applications in resource-constrained settings [3]. Rather than processing the whole facial image in a single shot, the model divides fixed-size overlapping patches and employs an attention mechanism to construct informative local representations along the patches [3]. By this design, PAtt-Lite is more robust to occlusion and pose variation by concentrating on the most informative parts of the face. The model attains a level of accuracy of up to 72.6% on the FER2013 dataset and only uses 9.1 million parameters with an inference time of 11.5 milliseconds on the Jetson Nano [3]. Along with its operational efficiency, PAtt-Lite also facilitates improved interpretability via attention heatmaps that visually identify those regions most dominant in driving classification decisions. These qualities make it well-suited to edge deployment situations in which both expedient inference and visual explanation are required. Owing to its balance of speed, accuracy, and architectural simplicity, we chose PAtt-Lite as the base model for the FER component of our multimodal system [3].

Sentiment Analysis The purpose of sentiment analysis is to extract opinions from text [4]. Bashiri and Naderi did a survey of the field of sentiment analysis, specifically focusing on comparing current transformer models, which show the most promise in the field at this time [4]. They start by outlining the unique problems posed by sentiment analysis [4]. The English language can be highly subjective, and speakers often do not use explicitly emotional language [4]. Nuances like sarcasm, irony, and polysemy (where two words have a different meaning) can make emotional sentiment more difficult to predict [4]. They also discuss more traditional methods that have been used for sentiment analysis [4]. Today, the field is advanced through work in the field of natural language processing, specifically transformers [4]. Bashiri and Naderi compare many different transformers according to their F1-score across 22 different datasets [4]. They find that T5 has the highest accuracy of the compared transformers [4]. They do not consider inference time as part of their evaluation metrics. They provide many of their datasets with the paper, and we have selected a few of these datasets to use for training our model [4].

Devlin et al. had a major breakthrough in pre-training of transformers with their framework BERT [5]. They leverage that idea that context in sentences moves both ways: words at the beginning of a sentence inform the meaning of words at the end of a sentence, and words at the end of the sentence inform the meaning of words at the beginning [5]. Pre-training strategies had previously focused on unidirectional context, as it is not possible to train a model on the left-to-right and right-to-left context on the same sentence, as it will have already seen the data by the time it is read from the opposite direction [5]. Instead, they use two pre-training strategies to learn bidirectional context: Masked Language Model (MLM) and Next Sentence Prediction [5]. MLM randomly masks words in a sentence and asks the model to guess what the word is, ensuring that the words both to the left and right of a word are considered in the prediction [5]. Next sentence prediction gives the model two sentences and asks does sentence B directly follow sentence A [5]. Devlin et al. showed that BERT achieved state-of-the-art performance on a variety of language tasks, demonstrating that it is not task-dependent [5].

In [14], the authors use the raw audio signal for Speech Emotion Recognition (SER). They claim the extra information from the audio signal, such as pitch and pause duration, provides critical insights into the speaker's emotions that would be lost with a speech-to-text translation [14]. The conventional approach to SER is to use Convolutional Neural Networks (CNNs) to extract features from a preprocessed audio signal (including signal transformations like Fast Fourier transformation) [14]. However, the authors in [14] argue that a multimodal approach that doesn't remove this temporal information is more equipped for SER. For their approach, they use a Spiking Neural Network (SNN) to create spike trains for each emotion that take these other speech features into account [14]. They did show promising results, achieving an accuracy of 65.3% and beating the current state-of-the-art for SER [14].

Multimodal Data Fusion The purpose of fusing independent data modalities is to add context to situations whose classifications are not one-dimensional. Zhao et al. propose multimodal sentiment analysis techniques that combine text from BERT and facial expression classifications from DINOv2 [6]. They compare fusion strategies including Basic, Self-Attention, and Dual Attention Fusion. The basic fusion model surprisingly outperforms the others, with up to 73% accuracy, compared to 53% from unimodal inputs. The authors note that limitations in dataset annotation quality, especially in Memotion, can lead to inaccuracies, but their results still surpass previous state-of-the-art.

A deep CNN with late fusion technique was performed by Dixit et al., using lightweight CNNs to prioritize efficiency in real-time environments [7]. Their model fused four modalities: text, audio, image, and video. Trained on the CMU MOSEI dataset, which intentionally includes class imbalance, the model achieved 85.85% accuracy and 83.0 F1-score. They note the benefits of data augmentation to handle imbalance and suggest future improvements through transfer learning. Their work demonstrates that multimodal fusion has real potential in areas like mental health, behavior monitoring, and political analysis.

III. Proposed Work

A. *Facial Expression Recognition*

The facial expression recognition component of our system uses a custom implementation of the PAtt-Lite (Patch Attention Lite) architecture, specifically designed for efficient deployment on edge devices [3]. PAtt-Lite combines the efficiency of MobileNetV1 with a self-attention mechanism to achieve high accuracy while maintaining a small model footprint suitable for deployment on our Raspberry Pi 5 target platform.

The model architecture consists of three key components:

1) **Backbone Network**: We utilize a pre-trained MobileNetV1 CNN as the primary feature extractor, truncated at layer -29 to reduce computational complexity while preserving important feature extraction capabilities. The backbone is kept frozen during training to leverage the pre-trained weights for efficient training and to prevent overfitting on our relatively small dataset.
2) **Patch Extraction Block**: After the backbone extracts global features, our patch extraction module processes

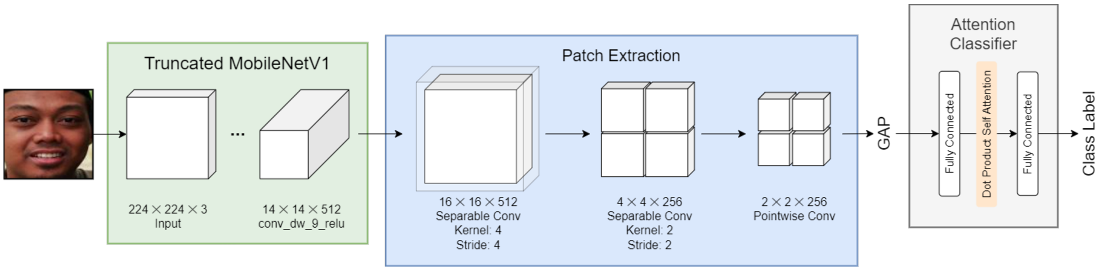

Fig. 1. Architecture of the PAtt-Lite model showing the three main components: MobileNetV1 backbone for feature extraction, Patch Extraction block for local feature processing, and Self-Attention Classification Module for emotion recognition.

these features to capture local facial details at different scales [3]. The module employs three convolutional layers:

- A separable convolution with kernel size 4, stride 4, and 256 filters to capture larger facial regions
- A second separable convolution with kernel size 2, stride 2, and 256 filters for medium-scale features
- A pointwise convolution (kernel size 1) with 256 filters to combine features across channels

3) **Self-Attention and Classification Module**: Following global average pooling, features are processed through a self-attention mechanism that helps the model focus on the most emotionally relevant facial features [3]. This is followed by dropout for regularization, a pre-classification dense layer with batch normalization, and a final softmax layer that outputs probabilities across seven emotion classes.

For our multimodal fusion approach, we map the seven emotion classes to a simplified positive/negative/neutral sentiment scale to align with the sentiment analysis model.

B. Sentiment Analysis

The sentiment analysis model is a BERT architecture with a sentiment analysis binary classification head [5]. Binary classification was chosen over multi-label with consideration for efficiency in resource constrained environments. The underlying model is the pretrained ”bert-base-uncased” released by Google through Hugging Face [8]. The model is 110 million parameters [8]. The model configuration and tokenizer were provided by Hugging Face [9].

Fine-tuning. The classification head followed the BERT model, consisting of a dropout layer with a 30% probability and a linear layer with binary output representing positive or negative sentiment. This is adapted from the Hugging Face implementation ”BertForSequenceClassification” [9]. The hyperparameters are set up as they are in [5]: an Adam optimizer is used with a learning rate of 1e-5, and beta coefficients of 0.9, 0.999 and an epsilon of 1e-8 are used, as is default for PyTorch’s Adam optimizer implementation.

Data. Three datasets provided by [4] are used: Archeage, Ntua, and HCR. These datasets were relatively small, so the IMDB movie review dataset was also included [10]. This brought the combined dataset up to almost 54,000 reviews labeled as either positive or negative. The class distribution of positive and negative reviews is split almost evenly. The combined dataset was split with 70% being used for training, 20% used for validation, and 10% used as the test dataset.

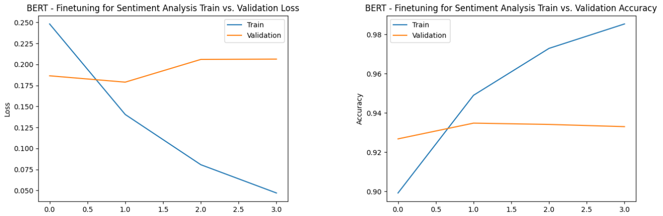

Fig. 2. Plots of the average loss (left) and accuracy (right) results after every epoch for the training and validation datasets.

Training. A batch size of 16 was used to train for four epochs. Shortly after the second epoch, the loss stopped decreasing and that best model was used in our application. The loss and accuracy curves in Figure 2 show that just two epochs were needed to reach a high accuracy. The best model received an average accuracy of 93.01% on the test dataset.

C. Multimodal Data Fusion

The purpose of fusing independent data modalities is to add context to situations whose classifications are not one-dimensional, e.g. facial expression analysis, medical diagnoses.

In research done by T. Zhao et al., multimodal sentiment analysis techniques, using image and text, are combined. The textual information gathered from Bidirectional Encoder Representations from Transformers (BERT) and facial expression classifications derived from the DINOv2 visual transformer [6]. Three fusion strategies are compared - Basic, Self-Attention, and Dual Attention Fusion. These strategies proposed by the authors yields equivalent-or-surpassing previous SOTA models in accuracy and F1 scores, with - interestingly - the Basic fusion model outperforming the proposed self- and dual-attention-based models [6]. Challenges include a potential inaccuracy in the annotations from the datasets (specifically the Memotion dataset) - the issue of oversimplifying human emotions to just ’positive’ ’neutral’ and ’negative’ leads to the aforementioned inaccuracy; for example, a human may be smiling in a negative manner - not accurately captured by the annotator or, consequently, the machine learning models [6]. Regardless, the multimodal fusion strategies each yield significantly improved results compared to unimodal classification (derived from text-only or visual-only), peaking at 73% accuracy compared to best-case 53% accuracy from the unimodal counterpart [6]. While there exists room for improvement, these findings suggest a promising future for multimodal fusion techniques for classification tasks.

A deep CNN with late fusion technique was performed by C. Dixit et al., using lightweight CNNs (as opposed to transformer models) to prioritize efficiency in real-time environments [7]. This approach focused on late-stage data fusion, and combined four separate modalities: text, audio, image, and video [7]. After each modality was trained and tested, a fused model was built and tested on a multi-modal human emotion data set containing six classes, ”CMU MOSEI dataset” [7]. Most notably about this dataset is the intentional imbalance in the classes - the original creators recognized that feelings of disgust were less prevalent in the real-world than positive emotions, so there is an intentional skewing to reflect that in training – according to C. Dixit et al., techniques such as data augmentation ensures no negative impact from the imbalance. The proposed model yields results surpassing SOTA recent models, with 85.85% and 83 for accuracy and F1-Score, respectively [7]. The author(s) suggests improvements could be made by incorporating transfer learning via pre-trained models and balancing the classes of emotions in the data set used [7]. Regardless, it is proven that the real-time application would be useful - for example, in the mental health sector (patient care), behavioral patterns of criminals to determine the possibility of repetition, analysis of political figures' behaviors, etc. [7].

For our multimodal emotion recognition system, we implement a late-decision fusion architecture that combines the outputs from both the PAtt-Lite facial expression recognition model and the BERT sentiment analysis model. This approach was selected based on the findings from Dixit et al. [7], which demonstrated the effectiveness of late fusion for real-time emotion analysis applications on resource-constrained devices.

Our late-decision fusion model consists of the following components:

1) **Input Mapping**: The model takes the classification results from the sentiment analysis (positive/negative) and facial expression recognition (seven emotion classes) models. The facial expressions are mapped to a simplified scale as shown in Table I.
2) **Tone-Based Weighting Mechanism**: We implement a weighting system that adjusts the influence of each modality based on confidence scores. This allows the system to rely more heavily on the more accurate sentiment analysis model when facial expressions are ambiguous.
3) **Feature Fusion**: The outputs from both models are concatenated to form a joint representation. This concatenation creates a 9-dimensional feature vector that captures both verbal and non-verbal emotional cues.
4) **Classification Head**: The final component consists of a fully-connected linear layer that reduces the 9-dimensional input features to 3 outputs, classifying the emotional state as ”positive,” ”neutral,” or ”negative.” A softmax activation function is applied to produce the final classification probabilities.

This fusion architecture enables our system to leverage the complementary strengths of both modalities, providing a more robust emotion recognition capability than either modality alone could achieve. The lightweight design of the fusion component ensures minimal additional computational overhead, maintaining real-time performance on our Raspberry Pi 5 target platform.

IV. Simulation Results

A. Deployment

The project is deployed to a Raspberry Pi 5. A camera and a microphone are used to collect information from a live participant and generate a resulting emotion. The image and speech samples are collected at the same time. The speech is then translated to text. The inputs then pass through the facial expression recognition and sentiment analysis models as shown in the diagram below. The models output results which are finally combined by the late fusion model.

B. Facial Expression Recognition Results

The PAtt-Lite model was trained on the FER2013 dataset, which contains 35,887 grayscale facial images at 48×48 pixel resolution labeled with seven emotion categories. We preprocessed the dataset by converting grayscale images to RGB by replicating the channel and resizing images to 120×120 pixels for consistent input to our model.

The model was trained using the Adam optimizer with a learning rate of 1e-3. Data augmentation with random horizontal flips and contrast adjustments was implemented to improve generalization performance. While the original PAtt-Lite implementation reported achieving up to 92.5% accuracy on the FER2013 dataset, our implementation achieved 62.15% accuracy on the test set [3]. This discrepancy can be attributed to several factors:

- The original PAtt-Lite was trained on a combination of datasets including AffectNet, which contains more diverse and higher quality facial expression samples than FER2013 alone [3]
- Our implementation used a simplified training pipeline with fewer optimization techniques
- The limited computational resources available on our Raspberry Pi 5 required some architectural simplifications

For integration with the multimodal system, we implemented a mapping function that converts the seven-class emotion prediction to a simplified sentiment scale, as shown in Table I.

This mapping allows the facial expression predictions to be combined with text sentiment analysis in our late fusion model. Despite the lower accuracy compared to the original

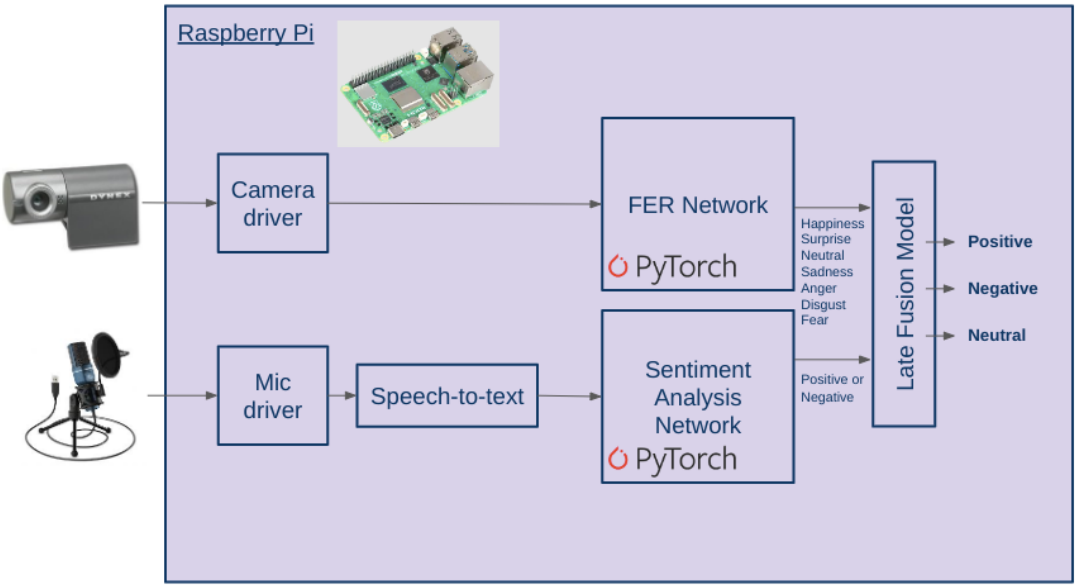

Fig. 3. The input and output flow of the project deployed on the Raspberry Pi, with images from [11], [12], and [13].

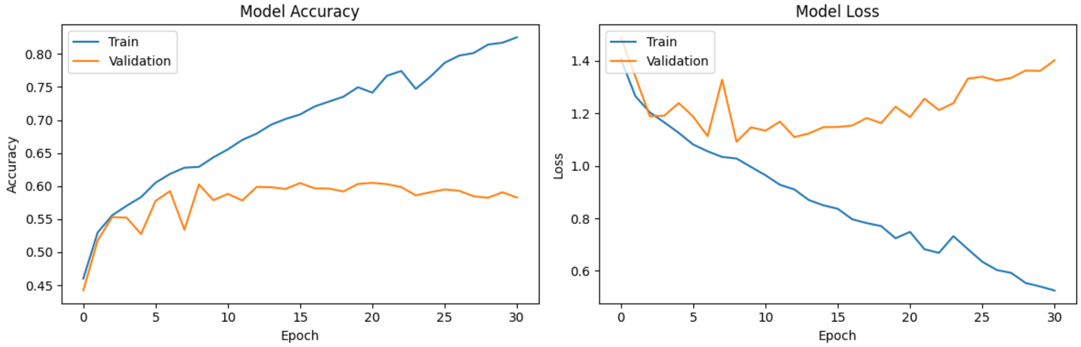

Fig. 4. Training accuracy (left) and loss (right) for our PAtt-Lite implementation showing model convergence around 30 epochs. The validation accuracy plateaus around 60%, showing potential for further optimization.

TABLE I
MAPPING FROM FER EMOTIONS TO SENTIMENT CATEGORIES

Emotion	Sentiment Value
Happiness	Positive (+1)
Neutral	Neutral (0)
Surprise	Neutral (0)
Anger	Negative (-1)
Disgust	Negative (-1)
Fear	Negative (-1)
Sadness	Negative (-1)

implementation, our model still provides useful emotional context that complements the sentiment analysis when deployed on the Raspberry Pi 5.

When evaluated on the challenging subsets (occlusion and posed faces), our FER model showed particular difficulty with the recognition of disgust and fear expressions, which is consistent with patterns observed in the literature. These challenges further highlight the benefits of our multimodal approach, where text sentiment can provide additional context when visual cues are ambiguous.

C. Sentiment Analysis

The sentiment analysis model is deployed to the Raspberry Pi CPU. The inference speed was fast enough for our application at approximately 313 milliseconds. However, as the application expands to include a participant continually speaking at a faster pace, a faster inference speed is desirable. Potential improvements for future work are discussed in Section 5.

The high accuracy obtained on the test dataset was reflected in the performance in the deployed application. Table 2 shows output from the sentiment analysis module alone on the Raspberry Pi.

It is shown that inputs with emotionally charged language are predicted as expected. However, the last two inputs are provided in an attempt to ”trick” the model, showing that text sentiment analysis alone is not capable of recognizing sarcasm and struggles with common idioms. These examples illustrate

TABLE II
THE INPUT (LEFT), EXPECTED OUTPUT (MIDDLE) AND ACTUAL OUTPUT (RIGHT) OF THE SENTIMENT ANALYSIS MODEL DEPLOYED ON THE RASPBERRY PI.

Input	Actual Emotion	Predicted Emotion
”the weather is beautiful today”	Positive	Positive
”i’m so disappointed”	Negative	Negative
”i love you”	Positive	Positive
”this is the worst”	Negative	Negative
”great! this is just what i needed today”	Negative	Positive
”it’s raining cats and dogs”	Negative	Positive

the need for a multimodal approach that considers other cues that would be necessary for human-to-human emotion recognition. Unfortunately, we were not able to prove this hypothesis. This is further discussed in the conclusions and future work of section 5.

V. CONCLUSION AND FUTURE WORK

We have created a framework for a multimodal approach to emotion recognition and deployed it to an edge device. Unfortunately, the combined accuracy of the models does not support our hypothesis that multimodal emotion recognition will be more accurate than single-modal emotion recognition. However, we believe this is due to the low accuracy of the facial recognition model and ineffective fusion of the models. With future improvements, we believe our hypothesis will be supported.

Our system demonstrated significant strengths in text sentiment analysis (93.01% accuracy) but faced challenges in facial expression recognition (62.15% accuracy). We observed that the late fusion model sometimes failed to properly weight the more accurate sentiment analysis prediction when combining results. Nevertheless, the successful deployment on the Raspberry Pi 5 with reasonable inference times (approximately 313 milliseconds for sentiment analysis and 1.5 seconds for FER) demonstrates the feasibility of edge-based emotion recognition for privacy-sensitive applications.

Future work could include several improvements to address the limitations we identified:

- **Improved Facial Expression Recognition:** Enhancing the FER model’s accuracy could be achieved through different training strategies such as progressive learning, curriculum learning, or using advanced data augmentation techniques in addition to training on more balanced datasets. Implementing these approaches would help the model better handle challenging cases like occlusion and extreme pose variations.
- **Integrated Multimodal Datasets:** Rather than using separate datasets for facial expression and sentiment analysis, future work should utilize datasets like CMU-MOSEI that contain synchronized facial expressions and speech from the same subjects. This would enable better alignment between modalities and more effective multimodal learning.
- **BERT Optimization:** The sentiment analysis model inference speed could be improved through model quantization, which reduces numerical precision while maintaining performance. Alternatively, we could implement DistilBERT, a smaller model that contains distilled knowledge from a trained BERT model, significantly reducing inference time without substantial accuracy loss.
- **Tri-modal Approach:** As demonstrated in [14], incorporating speech acoustic features alongside facial expressions and text sentiment could provide a more complete representation of emotion. This approach would capture prosodic elements like tone, pitch, and speaking rate that carry significant emotional content.
- **Demo Improvements:** Our current demonstration system encountered significant practical challenges. The poor microphone and camera quality, combined with a speech-to-text module that struggled to understand speech in non-ideal conditions, resulted in participants needing to repeat their statements multiple times with minimal background noise before the speech could be processed. Improving these hardware components and implementing a more robust speech recognition system would enhance the system’s usability.
- **Hardware Acceleration:** Deploying the facial expression recognition model to a dedicated TPU (Tensor Processing Unit) could significantly improve inference speed. This would enable real-time processing of both modalities simultaneously.
- **Enhanced Fusion Techniques:** Instead of the current hard-coded, late-decision model, a trained fusion model would be more effective. This model could learn optimal weighting between modalities based on their reliability in different contexts. Ideally, it would be trained on a dataset containing ground truths for all modalities.
- **Improved Output Representation:** Future versions could report probabilities for each emotion class instead of single classifications, providing more nuanced emotional analysis. Additional metrics like uncertainty estimation would provide valuable context about prediction confidence.
- **Emotion-Focused Datasets:** The sentiment analysis component could benefit from datasets specifically designed for emotional content rather than product reviews. Such datasets would better capture the nuances of human emotional expression in natural conversation.

These improvements would address the key limitations we encountered and potentially validate our hypothesis that multimodal emotion recognition can outperform single-modal approaches. The integration of these enhancements would move the system closer to human-level emotion recognition capability, making it more valuable for applications in healthcare, customer service, and human-computer interaction.

VI. Summary of Contributions

Moorehead, J. implemented the facial expression recognition component of the multimodal system. This included researching lightweight model architectures suitable for edge deployment, implementing the PAtt-Lite architecture, and developing the training pipeline for the FER2013 dataset. Moorehead, J. also created the camera interface for capturing facial expressions and integrated the FER component with the overall system. Additionally, Moorehead, J. contributed to the system evaluation and conducted experiments to measure model accuracy and inference speed on the Raspberry Pi 5.

Willis, R. implemented the sentiment analysis component of the multimodal system. This involved selecting appropriate datasets, fine-tuning the BERT model for sentiment analysis, and optimizing the model for deployment on the Raspberry Pi 5. Willis, R. also developed the speech-to-text pipeline that converts audio input to text for sentiment analysis and conducted testing to evaluate the model's performance on sarcastic and idiomatic expressions.

Lind, H. did literature review on current multimodal fusion techniques, the advantages to fusing at different levels, and decided on the late-decision fusion strategy for the sentiment categorization. Lind, H. researched techniques for machine learning and pre-training, such as freezing layers of pre-trained models, for potential application to BERT and Patt-LITE. Lind, H. programmed the late-decision data fusion script/class that joined the results from Patt-LITE and BERT models, creating a now multi-faceted categorization script.

VII. Code Availability

The complete source code for this project is available on GitHub at: https://github.com/haley-dev-1/sentiment-analysis-CSCE790. The repository includes all model implementations, training scripts, and deployment code for the Raspberry Pi 5.

References

[1] F. S. Alreshidi et al., "A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines," Alexandria Engineering Journal, vol. 61, no. 12, pp. 12333–12361, 2022.

[2] S. Patel et al., "Light-FER: A Lightweight Facial Emotion Recognition System on Edge Devices," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 5, 2023.

[3] S. Saadi, D. Bolya, and A. Sinha, "PAtt-Lite: Lightweight Patch Attention for Efficient FER," preprint, 2024. Available: https://arxiv.org/abs/2401.08918

[4] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis," Knowl Inf Syst, vol. 66, pp. 7305–7361, 2024.

[5] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019.

[6] T. Zhao et al., "Enhancing sentiment analysis through Multimodal Fusion: A Bert-DINOv2 approach," arXiv:2503.07943, 2025.

[7] C. Dixit et al., "Deep CNN with late fusion for real time multimodal emotion recognition," Expert Systems with Applications, vol. 240, 2024, Art no. 122579.

[8] "BERT," Hugging Face. https://huggingface.co/collections/google/bert-release-64ff5e7a4be99045d1896dbc

[9] "BERT," Hugging Face Documentation. https://huggingface.co/docs/transformers/en/model_doc/bert

[10] A. Maas, "Large Movie Review Dataset." http://ai.stanford.edu/ amaas/data/sentiment/

[11] "Raspberry Pi 5," Raspberry Pi. https://www.raspberrypi.com/products/raspberry-pi-5/

[12] "Dynex DX-WEB1C 1.3MP Webcam," Amazon. https://www.amazon.com/Dynex-DX-WEB1C-1-3MP-Webcam/dp/B001AO2Q5W

[13] "TONOR TC-777 USB Microphone," TONOR. https://www.tonormic.com/products/tonor-tc-777-usb-microphone

[14] U. Singh, K. Abhishek, and H. K. Azad, "A Survey of Cutting-edge Multimodal Sentiment Analysis," Association for Computing Machinery, vol. 56, no. 9, 2024. https://doi.org/10.1145/3652149