HomeWork 2 Report

Multimodal machine-learning tasks on multiple emotion categories using features obtained from pre-trained models

NAME: Kavya Gengedulla Yeranarasappa
USC ID: 3206274199

1) Multimodal Emotion Classification with Mean Pooling

After extracting data and relevant features from the IEMOCAP dataset, mean pooling was performed on audio and visual data numpy arrays to standardize them. Subsequently, classification was conducted on four emotion classes: anger(0), sadness(1), happiness(2), and neutral(3). **SVC** was utilized for video and text modalities, while a **random forest classifier** was employed for audio. Hyperparameters were defined and optimized through grid search with 5-fold cross-validation.

Hyper parameters defined as below:

```
# Define hyperparameter grids for each classifier
param_grid_visual = {
    'C': [0.1, 1, 10],
    'kernel': ['linear', 'rbf', 'poly']
}

param_grid_audio = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20]
}

param_grid_text = {
    'C': [0.1, 1, 10],
    'kernel': ['linear', 'rbf', 'poly']
}
```

The resulting F1-micro scores were as follows:

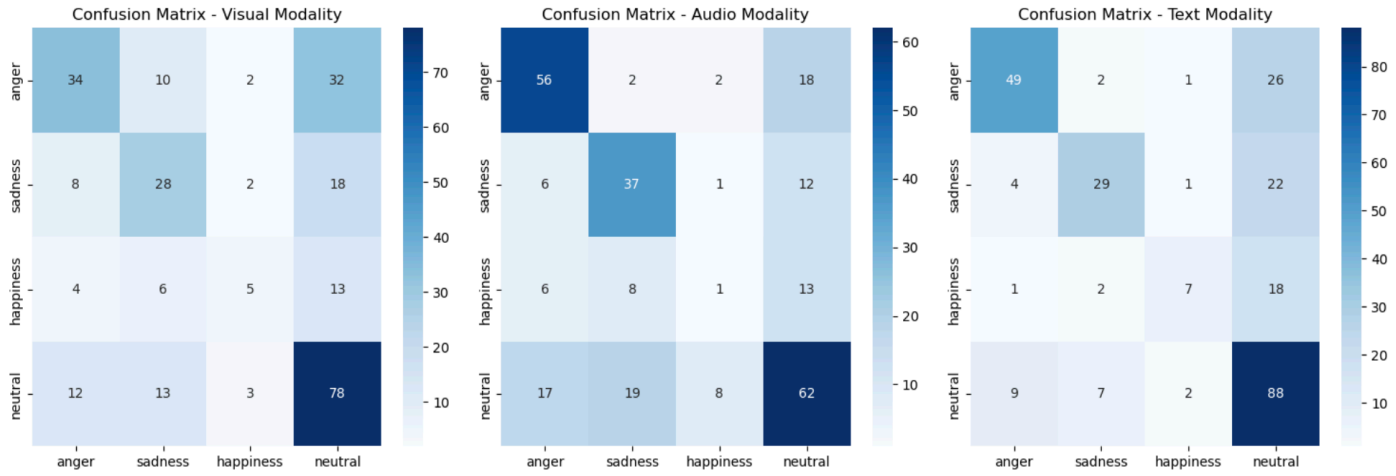
- F1-micro for Visual Modality: **0.4774**
- F1-micro for Audio Modality: **0.5857**
- F1-micro for Text Modality: **0.4851**

Overall, the random forest classifier performed best on the audio modality, achieving the highest F1-micro score. This suggests that the audio features were more discriminative for the classification task compared to visual and textual features.

2) Handling Class Imbalance and Confusion Matrix

To address the class imbalance issue, various techniques like **SMOTE** were explored, but ultimately, **RandomOverSampler** from the imbalanced-learn library was selected for its superior performance. This method effectively handled class imbalances which oversampled the minority class through duplication of samples until a balanced distribution was achieved.

The below confusion matrices visually represent the classification performance of each modality across the give four emotion classes:



3) Early and Late fusion

Early fusion was achieved by concatenating the features from different modalities (visual, audio, and text) into a single feature vector at the input level. This concatenated feature vector was then used to train classifiers that were initially defined.
Late fusion, on the other hand, involved obtaining predictions from unimodal classifiers for each modality separately. These predictions were then combined using a majority vote mechanism over the outputs of the unimodal models to make decisions.

Below were the accuracies obtained:

- Accuracy for Early Fusion: **69.40%**
- Accuracy for Late Fusion: **67.16%**

Observation of fusion results:
Early fusion demonstrates a slight advantage over late fusion in terms of accuracy. This suggests that directly combining features from various modalities at the feature level yields slightly better classification performance compared to aggregating predictions from individual modality classifiers. However, the discrepancy in accuracy between the two fusion methods is marginal.

4) Interpretation of unimodal and multimodal classification

The results indicate that in unimodal classification, the audio modality achieved the highest F1 score (0.57), followed by text (0.48) and visual (0.477) modalities. However, in multimodal classification, particularly with early fusion, performance surpassed that of unimodal classifiers. This suggests that integrating information from multiple modalities early in the process provides better classification accuracy. Early fusion likely benefited from the diverse information offered by different modalities, resulting in improved performance compared to late fusion. Although the difference in accuracy between early and late fusion is small, the advantage of integrating multiple modalities for enhanced classification performance is evident, particularly with early fusion.

The results clearly show that **multimodal classification performed better** when compared to unimodal classification.