**Results**

1. **Dataset**

Our dataset used in this work includes audio, text and video data of response to daily questions from 417 students aged 13-15. The data was collected in a junior school in Guangzhou. At the same time, we asked these students to finish a psychological questionnaire. The results of questionnaire were regarded as the gold standard to measure each students’ mental health. The results showed that there were 322 healthy samples and 95 unhealthy samples among these students.

1. **Results of Uni-modal Classification**

As a baseline, we used SVM trained on features extracted from each modality separately. To ensure the reliability of the results, we run the same model with different random seed for 10 times and report the maximum and the average accuracy and F1 score of the models (the average value in brackets) (we have adopted this approach in all experiments mentioned in this paper). The results are shown in Table 1.

The table shows that the best results were obtained for textual modality; the visual modality performed worse and the audio was the worst one. There is about 4% gap between the best and the worst result which is acceptable.

|  |  |  |  |
| --- | --- | --- | --- |
| **Modality** | **Dimension of feature** | **ACC (%)** | **F1 (%)** |
| Text | 768 | **68.75 (64.38)** | **76.92 (67.38)** |
| Visual | 768 | 66.66 (57.92) | 70.59 (59.83) |
| Audio | 512 | 64.58 (56.67) | 69.84 (61.71) |

Table 1. Results of uni-modal classification.

1. **Results with Model-agnostic Fusion**

We tried two kinds of model-agnostic approaches of feature fusion. One is early fusion in which the features extracted from the three modalities are simply concatenated together and used to train an SVM classifier. The other method is late fusion. Three SVM classifiers were trained on each modality and their decisions were combined together to train a final SVM classifier.

As Table 2. Reports, the best accuracy of late fusion is better than that of early fusion but the average accuracy and F1 score of early fusion is better. The performance of both fusion strategies outperformed unimodal classification mentioned in the last section.

1. **Results with Model-based Fusion**

Four model-based methods were tested in our experiments: MKL (Multiple Kernel Learning), TEN (Tensor Fusion Network), LMF (Low-rank Multimodal Fusion) and EmbraceNet.

We used the same MKL implementation as in [1] which is designed to deal with heterogeneous data. In TEN and LMF, the features of the textual, audio and video modality were first fed into a sub net and the dimension of features were changed from (768, 512, 768) to (16, 16, 64) before fusion. In EmbraceNet, the Embracement Size (dimension of the fused feature) was set to 256.

The performance of MKL is similar with that of early fusion and late fusion, while TEN and LMF is worse than early fusion, late fusion and MKL. This may be because the feature dimension obtained after fusion is very high.

EmbraceNet obtained the best results, which is about 10% higher than LMF.

|  |  |  |
| --- | --- | --- |
| **Method** | **ACC (%)** | **F1 (%)** |
| Early fusion + SVM | 77.08 (62.30) | 80.00 (63.69) |
| Late fusion + SVM | 79.17 (61.19) | 80.77 (57.78) |
| MKL | 79.17 (60.96) | 79.17 (61.45) |
| TEN | 72.92 (63.75) | 74.07 (63.83) |
| LMF | 75.00 (61.17) | 69.57 (51.59) |
| EmbraceNet | **85.42 (68.89)** | 86.27 (68.55) |

Table 2. Results of fusion methods

**Conclusions & Discussion**

**Reference**

[1] Poria, S., Cambria, E., & Gelbukh, A. (2015, September). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 2539-2544).