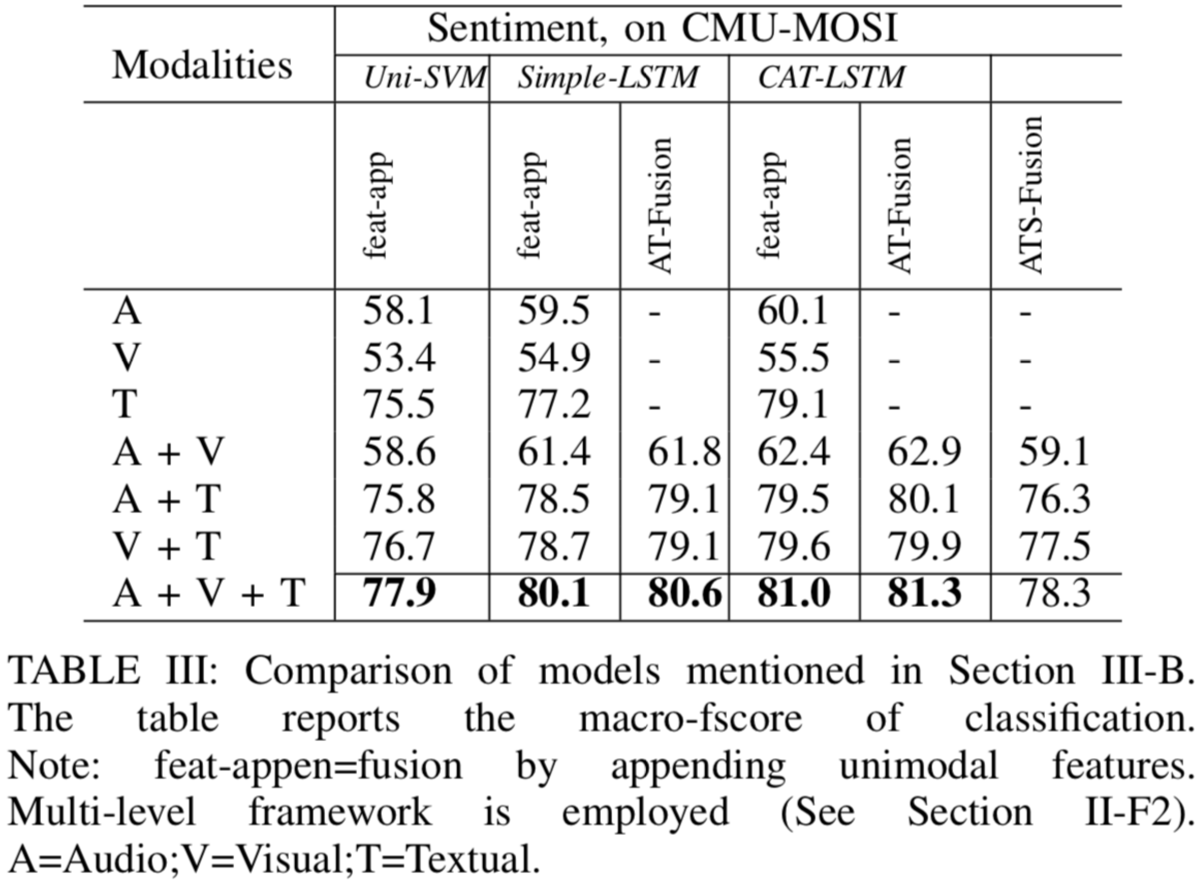
Audio Sentiment Analysis Experiment Results

MOSI Dataset

Method	2-class	3-class	5-class	7-class
SOTA(%)	60.3	N/A	30.7	34.7
Our Method(%)	68.6	49.7	34.2	26.8

Reference



Modalities	Sentiment, on CMU-MOSI					
	Uni-SVM	*Simple-LSTM*		*CAT-LSTM*		
	feat-app	feat-app	AT-Fusion	feat-app	AT-Fusion	ATS-Fusion
A	58.1	59.5	-	60.1	-	-
V	53.4	54.9	-	55.5	-	-
T	75.5	77.2	-	79.1	-	-
A + V	58.6	61.4	61.8	62.4	62.9	59.1
A + T	75.8	78.5	79.1	79.5	80.1	76.3
V + T	76.7	78.7	79.1	79.6	79.9	77.5
A + V + T	**77.9**	**80.1**	**80.6**	**81.0**	**81.3**	78.3

TABLE III: Comparison of models mentioned in Section III-B. The table reports the macro-fscore of classification. Note: feat-appen=fusion by appending unimodal features. Multi-level framework is employed (See Section II-F2). A=Audio;V=Visual;T=Textual.

Modality	MOSI				
	hierarchical (%)				non-hier (%)
	uni-SVM	h-LSTM	sc-LSTM	bc-LSTM	
T	75.5	77.4	77.6	**78.1**	
V	53.1	55.2	55.6	**55.8**	
A	58.5	59.6	59.9	**60.3**	
T + V	76.7	78.9	79.9	**80.2**	78.5
T + A	75.8	78.3	78.8	**79.3**	78.2
V + A	58.6	61.5	61.8	**62.1**	60.3
T + V + A	77.9	78.1	78.6	**80.3**	78.1

Acoustic Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
HL-RNN	63.4	64.2	25.9	**1.21**	0.34
Adieu-Net	59.2	60.6	25.1	1.29	0.31
SER-LSTM	55.4	56.1	24.2	1.36	0.23
CMKL-A	52.6	58.5	**29.1**	-	-
SAL-CNN-A	62.1	-	-	-	-
SVM-MD-A	56.3	58.0	24.6	1.29	0.28
TFN$_{acoustic}$	**65.1**	**67.3**	27.5	1.23	**0.36**
$\Delta^{SOTA}_{acoustic}$	$\uparrow$ 1.7	$\uparrow$ 3.1	$\downarrow$ 1.6	$\uparrow$ 0.02	$\uparrow$ 0.02

Table 5: Acoustic Sentiment Analysis. Comparison with state-of-the-art approaches for audio sentiment analysis and emotion recognition. $\Delta^{SOTA}_{acoustic}$ shows improvement.

	Binary		Multiclass	Regression	
Method	A^2	F1	A^7	MAE	Corr
Majority	50.2	50.1	17.5	1.864	0.057
RF	56.4	56.3	21.3	-	-
SVM-MD	71.6	72.3	26.5	1.100	0.559
THMM	50.7	45.4	17.8	-	-
SAL-CNN	73.0	-	-	-	-
C-MKL	72.3	72.0	30.2	-	-
EF-HCRF$_{(\star)}$	65.3$_{(\text{h})}$	65.4$_{(\text{h})}$	24.6$_{(\text{l})}$	-	-
MV-HCRF$_{(\star)}$	65.6$_{(\text{s})}$	65.7$_{(\text{s})}$	24.6$_{(\text{l})}$	-	-
DF	72.3	72.1	26.8	1.143	0.518
EF-LSTM$_{(\star)}$	73.3$_{(\text{sb})}$	73.2$_{(\text{sb})}$	32.4$_{(\text{-})}$	1.023$_{(\text{-})}$	0.622$_{(\text{-})}$
MV-LSTM	73.9	74.0	33.2	1.019	0.601
BC-LSTM	73.9	73.9	28.7	1.079	0.581
TFN	74.6	74.5	28.7	1.040	0.587
MARN (no MAB)	76.5	76.5	30.8	0.998	0.582
MARN (no $\mathcal{A}$)	59.3$_{(3)}$	36.0$_{(3)}$	22.0$_{(3)}$	1.438$_{(5)}$	0.060$_{(5)}$
MARN	**77.1**$_{(4)}$	**77.0**$_{(4)}$	**34.7**$_{(3)}$	**0.968**$_{(4)}$	**0.625**$_{(5)}$
Human	85.7	87.5	53.9	0.710	0.820

Table 1: Sentiment prediction results on CMU-MOSI test set using multimodal methods. Our model outperforms the previous baselines and the best scores are highlighted in bold.