MULTI-MODAL SENTIMENT ANALYSIS OF INTERVIEWS

Dataset 1: CMU-MOSEI for multi-modal analysis

- CMU Multi-modal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset.
- CMU-MOSEI dataset consists of 3,229 videos spanning over 23,000 utterances from
  more than 1,000 online YouTube speakers.
- The training, validation & test set comprises
  of 16216, 1835 & 4625 utterances, respectively.
- CMU-Multi-modal Data SDK1 for downloading and feature extraction.Dataset statistics for CMU-MOSEI
- The raw data was downloaded from : http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/

	TRAIN	VALIDATION	TEST
VIDEOS	2250	300	679
`UTTERANCE	16216	1835	4625

Dataset 1: Feature engineering

- The visual features were extracted from each segment of each utterance ID using Facets2 - 35 dimension vector
- The audio features extracted using CovaRep - 74 dimension vector
- Textual features using 300 embedding dimension Glove vector

Dataset 1: Supervised Learning multi-modal Bi-GRU on trimodal model(explained in detail later)

Dataset 2: MELD

- A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation has been accepted as a full paper at ACL 2019.
- MELD has more than 1400 dialogues and 13000 utterances from Friends TV series.
- Each utterance in a dialogue has been labeled by any of these seven emotions and sentiments (positive, negative and neutral)
- The raw data was downloaded from : http://web.eecs.umich.edu/~mihalcea/downloads/MELD.Raw.tar.gz

Train	Dev	Test	
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

Dataset 2: Feature engineering

- The audio features extracted using openSMILE and then followed by L2-based feature selection using SVM - 1611 feature vector for emotion classes and 1422 feature vector size for sentiment classes
- Textual features using 300 embedding dimension Glove vector

Dataset 2: Supervised learning Bi-LSTM network on bi-modal acoustic-text input(explained later)

First model stage trained on dataset 1

Supervised Learning attention-based multi-modal Bi-GRU on trimodal model

Code implementation : Stage1

Network Architecture

(inspired from https://www.aclweb.org/anthology/D18-1382)

Hyperparameters

Bi-GRU : 300 neurons & dropout = 0.3

Dropout for regularisation : 0.3

Optimiser : Adam ; Epochs : 5; Batch size : 20

Input : Concatenating three modalities

Output : 3 Sentiment classes(positive ,negative ,neutral)

Test Accuracy : 71%

Second model stage trained on dataset 2

Supervised Learning multi-modal Bi-LSTM on bimodal model

Code implementation : Stage 2

Hyperparameters

Bi-LSTM : 300 neurons & dropout = 0.3

Optimiser : Adam ; Epochs : 10 ; Batch size : 20

Input : Concatenating two modalities

Output : 3 Sentiment classes(positive ,negative ,neutral)

Trained on t2.xlarge AWS Sagemaker instance

Test Accuracy. : 67%