Domain Background

Real-world application:

1. There are many students who fear interview process.They get evaluated based on their confidence, their knowledge and their posture and many other latent factors. But they never learn what's wrong if they get rejected. So my main aim is to create an online portal which gives feedback on how they answered during the interview process by building highly accurate Multi-modal analysis of utterances.

Researches in this field:

1. https://www.researchgate.net/publication/330982553_Multimodal_Sentiment_Analysis_A_Survey_and_Comparison
2. https://www.aclweb.org/anthology/D18-1382

Problem Statement

1. Since sentiment of an utterance cannot be just defined from text independently or audio independently, hence there is a need to develop highly accurate multi modal models which consider text,image and speech.
2. Since model has to have a true negative rate higher than true positive for this problem to let the student that he has these negative sentiments to improve upon, recall and precision score matter.

Dataset

- 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

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

Solution Statement

Since all modalities cannot be equally weighted to predict a sentiment, the solution is to use weighted based approaches to give each modal a weight and then concat the weighted vectors to produce a positive/negative/neutral sentiment.The supervised learning approach is to create a feature vector on combinations of pairs of modalities (text-acoustic, acoustic-video,text-video) along with individual vectors.CMU Multi-Modal Data SDK will be used to extract features from the dataset.The textual, visual and acoustic features were extracted by GloVe , Facets2 & CovaRep, respectively.

Benchmark Model

https://www.aclweb.org/anthology/D18-1382
This paper gets 82.3% accuracy on MOSI dataset

Evaluation Metrics

Evaluation criteria is to measure the rate of false positives since the motive of the project is to accurately predict true negative sentiments and use the prediction to help people in interviews.

Project Design

Preprocessing
Text feature - of each utterance using Glove
Acoustic feature - CovaRep
Visual feature - Facets2

Network design