Unimodal Pipeline:

AUDIO CNN:

Training accuracy: 70%

Testing:

Test Accuracy: 57.64%

Precision (weighted): 0.6151

Recall (weighted): 0.5764

F1-Score (weighted): 0.5685

TEXT RNN:

Tried implementing the RNN network using LSTM after transcription using Whisper Open AI, got an accuracy of just 14%. The reason might be because of the length of text in each video, the average length I got was just 6 tokens because of which model was unable to learn.

MULTIMODAL:

Then tried implementing the Multimodal – Audio + Text, using the late fusion method. With a 80% weightage from Text RNN and 20% from Audio CNN, got an accuracy of just 20%.