Introduction of Problem Context

According to **(Picard),** emotions are a mental state where one experiences pleasure or displeasure with high intensity. They are a very necessary semantic component in human interaction - often, without the context of the speakers emotion, the intentions behind utterances can be ambiguous **(Automatic Emotion Recognition)**. The field of Affective computing aims to understand this phenomenon so that interactions between humans and machines become more naturalistic. Automatic Emotion Recognition has made significant strides in the previous decade but there are still many areas of unexplored territory. Up until the Audio/Visual Emotion Challenge in 2019, there were few works in affective computing recognition literature that supported the common idea that emotions conveyed by facial expressions are mostly universal across cultures **(AVEC 2018, 2019).** There still exist, however, some barriers to universal emotion recognition. It has been found that training machines to recognise emotion from similar language families have shown more accurate results. **(AVEC 2018).**

The dataset we use in this project are audio-visual recordings that have been collected “in the wild”. This phrase simply refers to the idea that standard webcams have been used for recording in a natural setting (home/ work place). The data is not preselected and the behaviour exhibited by the subjects are wholly spontaneous and naturalistic. **(AVEC 2019).** Literature has demonstrated great success with “in the lab” data where the variables are controlled, but this does not account for the noise present in real life situations that “in the wild” data can mimic.

Multiple modalities are exploited in this project meaning that both audio and video data shall be used to extract information about the emotional cues demonstrated. Whilst vocal expressions have been showed to provide fewer universal emotion markers than facial expressions **(AVEC 2019)**