abstract

August 19, 2022

Emotions are the windows to the nonverbal communication framework employed by the animal kingdom since inception. Humanity is not an exception to this practice. Often, throughout history, people have concentrated on the verbal aspect of the communication but recently, with the advances in Machine Learning and Artificial Intelligence, researchers have tried to delve into this realm to try detecting the emotions, their nuances and their role in effectively understanding communication.Teaching a machine that understands only bits, to gauge the emotions of a subject is the beauty of the modern Machine learning algorithms and frameworks. The technology is still at intermediate stages and needs further analysis and research on the same. This dissertation was aimed at evaluating the performance of the different emotion recognition systems across different modalities (facial expressions, audio and text).The study also analyzed how the emotion recognition systems performed in different contexts: posed vs spontaneous, ethnicity and gender. It also analyzed how the Emotion Recognition Systems(ERs) process the physical correlates or landmarks of the modalities (Action Units for face, F0,intensity, loudness, jitter,etc for audio). Similarly how the FERs process the action units were also analyzed. Both the Affectiva and Emotient reported dominant action units(AU) very similar to what has been reported in literature. Statistically significant differences were observed between the distribution of both the facial emotion recognition and speech systems.Differences were also observed in the system outputs when compared across categories like ethnicity and gender. The findings in this dissertation corroborated several studies on posed vs spontaneity, cross-culture variations. For the speech emotion recognition systems it was observed that the inter-annotator agreement rapidly decreased when the systems were trained on databases they have not seen before.For Automatic Speech Recognition (or ASR), it has been observed that Desript outperformed Google Cloud as the preferred service where GCP showed a Word Error Rate of about 80 percent. On the other hand Descript WER rate was around 28 percent. A 3-lexicon system was used to properly extract sentiments from the transcripts and the transcripts were also analyzed across four levels with a bag of words model. The dissertation also aims to introduce a preliminary model of multi-modal emotion detection and fusion as a part of future work