International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.188

I Know How You Are Feeling Now: A Multimodal Approach to Emotion Detection Using Transfer Learning

Dr Pratima Singh
Computer Science & Engineering
Ajay Kumar Garg Engineering College
Ghaziabad, India
Singhpratima@akgec.ac.in
Orchid ID:0000-0001-8937-8138

Ishaan Saxena
Computer Science & Engineering
Ajay Kumar Garg Engineering College
Ghaziabad, India
ishaan1910022@akgec.ac.in

Pranjali Singh
Computer Science & Engineering,
Banastali Vidyapith University,
Rajasthan, India
pranjalisingh2020@gmail.com

Dr Deepti Mishra
Computer Science & Engineering
Norwegian University of Science and Technology (NTNU)
Gjøvik, Innlandet, Norway
deepti.mishra@ntnu.no

Shreyas Maitrya
Electrical & Computer Engineering
Minneapolis,Minnesota,USA,
shreyasmaitreya34@gmail.com

Abstract **— Understanding Human has been a persistent pursuit of all Human being. Identifying human Emotions is the First step towards the same. We have developed a Transfer learning based Deep Neural Network model which works on multimodal data such as Text, Images, Video and Audio for the purpose of Emotion detection. Out of several emotional Models available in the Literature of psychology we have taken Plutchik wheels of Emotion as the base model to categorize the predicted emotions.**

Keywords — ***Emotion Detection, Plutchik Model, BERT, CNN, Multimodal data, Cognitive Psychology***

I. INTRODUCTION

Since time Immortal Understanding Human emotion has been a continued interest of human beings. In recent times there has been tremendous interest among researchers, to understand Human Emotions and their mind-sets at any point of time. There are several Emotional Models in the History of Neuropsychology or Cognitive Psychology. We have considered Plutchik wheels of Emotion due to its representation in the form of Wheel depicting Opposite Emotion on opposite arc of the wheel. In Plutchik Wheel Function Broad basis of classification can be Positive or Negative emotions, verbal or non-verbal, basic or complex [10]. Understanding of these emotions can only be done by observing Multimodal Data. Multimodal [11]data Include data collected from Audio, video, Text, Biomarkers, Brain Signals, Body gestures, Voice, Eye contact, Interpersonal distance(space) and many more. Thus true identification of human emotions is possible only through the study of Multimodal data.

Collecting the multimodal data, its storage, computation, & final Interpretation of the result of Combination of Multimodal data is another challenge, towards human emotion detection process. So usage of machine learning models with transfer learning based deep neural network seems to be one of the good solution to this problem [12,13].

Multi-modal data can present several challenges, including:

i) Integration: Combining multiple data sources and ensuring they are compatible and complement each other can be difficult.

Alignment: Ensuring that the different modalities are aligned in time and space can be a challenge.

ii) Missing data: Some modalities may not be available for certain instances, which can result in missing values and biases in the data.

iii) Heterogeneity: Different modalities can have different characteristics and distributions, which can lead to difficulties in modelling and representation.

iv) Interpretability: Models that integrate multiple modalities can be complex and difficult to interpret.

v) Annotation: Collecting and annotating multi-modal data can be time-consuming and expensive.

Multimodal approaches to emotion detection involve using multiple sources of data to detect emotions in an individual. This can include using information from audio, video, text, physiological signals, and other sources. The following are some common multimodal approaches used for emotion detection:

i) Audio-Visual Emotion Recognition: This approach combines audio and visual cues to detect emotions, such as facial expressions and vocal intonation.

ii) Physiological Emotion Detection: This approach uses physiological signals, such as heart rate and skin conductance, to detect emotions.

iii) Text-based Emotion Recognition: This approach uses natural language processing techniques to analyse text data, such as written or spoken words, to detect emotions.

iv) Hybrid Emotion Detection: This approach combines multiple modalities to achieve better accuracy in emotion detection.

v) These approaches have been used in a variety of applications, including human-computer interaction,

DOI: rp-9788770040723.188

affective computing, and psychology research. The choice of approach depends on the specific requirements of the application and the available data.

Off course Multimodal Data along with Transfer Learning Methods Appear with its own set of Problems of combining the results of Multimodal Data using Multiple Deep Learning Approach [1]. Over all approach to this Problem of Emotion Detection has led to the Following Research Questions:

RQ1). What are the Verbal and Non Verbal Indicators of Emotion Detection?

The Verbal indicators of emotion include:

1. Tone of voice: The pitch, volume, and inflection of a person's voice can convey different emotions. For example, a high-pitched, excited tone might indicate happiness, while a low, monotone voice might indicate sadness.
2. Word choice: The words a person chooses and the way they put them together can reveal their emotions. For example, someone might use more negative or angry language when they're feeling frustrated or upset.
3. Speech rate: The speed at which a person speaks can also convey emotions. For example, someone might speak quickly when they're excited or anxious, or slowly when they're sad or tired.
4. Pauses: A person's use of pauses in their speech can reveal their emotions. For example, someone might pause before speaking if they're feeling hesitant or uncertain.

Nonverbal indicators of emotion include:

1. Facial expressions: The way a person's face looks can reveal their emotions. For example, a smile usually indicates happiness, while a scowl usually indicates anger or frustration.
2. Body posture: The way a person holds their body can convey emotions. For example, someone who is happy or relaxed might stand or sit with good posture and open body language, while someone who is anxious or upset might slouch or cross their arms.
3. Gestures: The way a person moves their hands and arms can reveal their emotions. For example, someone who is excited might wave their arms around, while someone who is angry might clench their fists.
4. Eye contact: The way a person looks at others can convey emotions. For example, someone who is happy or confident might maintain strong eye contact, while someone who is anxious or uncertain might avoid eye contact.

It's important to note that these indicators can vary from person to person, and that it's often necessary to consider multiple indicators in order to accurately interpret someone's emotions. During our exploration we came across several theories of Psychology. Following are the Psychological Models Behind the Emotion Detection Model.

1) Shaver [15] out of 135 Emotional words they developed Abstract to concrete Emotional Hierarchy consisting of finally Synthesized to 6 emotions such as sadness, anger, fear, surprise joy, love.
2) OCC Model[16] : consist of 22 Emotion Type , Finally Merging & labelling into six Broad classes: They are the following six kinds of emotion groups and each of them includes several basic emotions: fortune-of-others (happy, resentment, gloating and pity), well-being (joy and distress), attribution (pride, shame, admiration, and reproach), attraction (love and hate), prospect relevance (satisfaction, fear, relief, and disappointment) and well-being/attribution compounds (gratification, remorse, gratitude, and anger)
3) Ekman's [17] emotion model very similar to Shaver's model, who claims to give more discrete measureable 6 classes of Emotions such as anger, fear, disgust, joy, sadness, and surprise
4) Alena [18] proposed measuring each emotion word into values of nine basic emotions: Anger, disgust, fear, guilt, interest, joy, sadness, shame, and surprise
5) Plutchik [29] Model has represented the collection of 32 micro emotions into 8 macro emotions based on Positive and Negative emotions. These are: Joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. These emotions have been represented in form of "Wheel of Emotions" where positive joy versus sadness and Negative emotions are placed on the opposite sides of the wheels.

\

Fig. 1. Plutchik's wheel of Emotions [29]

We are going to use Plutchik psychological model for our deep learning based emotion detection. This has been chosen because of the discreetness and placement of the positive & negative emotions on the opposite side of the wheel which help us identify the relative radian value of the emotions being considered.

For our current work we have considered following basic type of Data for the Multimodal Approach:

1) Text

DOI: rp-9788770040723.188

2) Audio

3) Video/ Images

RQ2) What are the common retrained model for different types of Emotion Detection Input

2.1 Text Based

1) For Text Input the such as RoBERTa ((Robustly Optimized BERT Approach), popular PreTrained Models are: BERT (Bidirectional Encoder Representations from Transformers) and its variants ALBERT (A Lite BERT).

BERT is a pre-trained language model developed by Google that can be fine-tuned for various natural language processing tasks, including emotion detection.

ALBERT is a pre-trained language model developed by researchers at Google that can be fine-tuned for various natural language processing tasks, including emotion detection.

RoBERTa: RoBERTa is a pre-trained language model developed by researchers at Facebook AI that can be fine-tuned for various natural language processing tasks, including emotion detection.

Other popular LLM (Large Language Model) based pre trained model for NLP based emotion detection are GPT2, ELMo, XLNet

GPT-2: GPT-2 is a pre-trained language model developed by OpenAI that can be fine-tuned for various natural language processing tasks, including emotion detection.

ELMo: ELMo is a pre-trained language model developed by researchers at the University of Washington that can be fine-tuned for various natural language processing tasks, including emotion detection.

XLNet: XLNet is a pre-trained language model developed by researchers at Carnegie Mellon University and Google that can be fine-tuned for various natural language processing tasks, including emotion detection.

2.2 For ECG Signals

There are several pre-trained models that have been developed to work with brain signals, such as electroencephalography (EEG) data, for tasks such as emotion detection, mental workload assessment, and diagnosis of neurological disorders. Here are a few examples:

i) DEAP: DEAP is a pre-trained model developed by researchers at the University of Toronto that can be used to classify emotions based on EEG data. The model was trained on a dataset of EEG recordings collected from 32 participants while they watched emotionally-evocative videos.

ii) BCI-DNN: BCI-DNN is a pre-trained model developed by researchers at the University of Freiburg that can be used to classify mental workload based on EEG data. The model was trained on a dataset of EEG recordings collected from 23 participants while they performed a visual search task under different levels of mental workload.

iii) E-LSTM: E-LSTM is a pre-trained model developed by researchers at the University of California, San Diego that can be used to classify neurological disorders based on EEG data. The model was trained on a dataset of EEG recordings collected from patients with various neurological disorders, including epilepsy, Alzheimer's disease, and Parkinson's disease.

2.3. For Audio or sound data

VGG-Voice: VGG-Voice is a pre-trained model developed by researchers at Oxford University that can be used to classify speaker identities based on voice data. The model was trained on a large dataset of voice recordings and has achieved high accuracy on speaker identification tasks.

i) Deep Speaker: Deep Speaker is a pre-trained model developed by researchers at Baidu that can be used to classify speaker identities based on voice data. The model was trained on a large dataset of voice recordings and has achieved high accuracy on speaker identification tasks.

ii) Audio Set: AudioSet is a pre-trained model developed by researchers at Google that can be used to classify audio events in a wide range of categories, including speech, music, and environmental sounds. The model was trained on a large dataset of audio recordings and has achieved good performance on various audio classification tasks.

iii) SoundNet: SoundNet is a pre-trained model developed by researchers at the University of Cambridge that can be used to classify audio events based on spectral features extracted from raw audio waveforms. The model was trained on a large dataset of audio recordings and has achieved good performance on various audio classification tasks.

2.4. Video/ Facial Expression

i) Facial Action Coding System (FACS) model and the DeepMoji model. Facial Expression Recognition (FER) models: These are pre-trained models that have been specifically developed to classify facial expressions in video data. Examples include the VGG-Face model mentioned earlier, as well as models such as FER2013 and FERPlus. These models typically use convolutional neural networks (CNNs) trained on large datasets of images and videos of facial expressions.

ii) Action Unit (AU) models: These are pre-trained models that have been specifically developed to classify facial action units (AUs) in video data. AUs are specific facial movements that can be used to identify and classify emotions. Examples of AU models include the Facial

iii) Video Classification models: There are also more general pre-trained models that have been developed for video classification tasks, which can potentially be fine-tuned for emotion detection. Examples include models such as 3D ResNets and I3D, which use 3D convolutional neural networks (CNNs) to classify video frames.

DOI: rp-9788770040723.188

RQ:3 common data set needed for handling multimodal data for emotion detection

There are several common datasets that can be used for handling multimodal data for emotion detection. Here are a few examples:

1. AFEW: The Acted Facial Expressions in the Wild (AFEW) dataset is a collection of videos of facial expressions in naturalistic settings. The dataset includes annotations of facial action units (AUs) and basic emotions.
2. EmotionX: The Emotion dataset is a collection of audio, text, and video data annotated with emotions and sentiments. The dataset includes a variety of data sources, such as social media posts, customer service conversations, and movie scripts.
3. MOSI: The Multimodal Opinion Sentiment and Influence (MOSI) dataset is a collection of video and audio data annotated with sentiments and opinions. The dataset includes a variety of data sources, such as political speeches, news articles, and movie reviews.
4. IEMOCAP: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset is a collection of audio, video, and text data annotated with emotions and sentiments. The dataset includes dyadic conversations between two actors in a variety of settings, such as customer service and therapy sessions.

RQ4: Which Approach should we use for Combining the Multimodal Models

There are various approached to Combine the Multimodal Data.

There are several ways to combine multimodal data, depending on the type of data and the task at hand. Some common techniques include:

i) Feature concatenation: This involves concatenating the feature vectors of the different modalities and feeding the resulting concatenated vector into a machine learning model. This is a simple and straightforward approach, but it assumes that the different modalities are independently informative and that their feature spaces have the same dimensionality.

ii) Feature fusion: This involves combining the feature vectors of the different modalities at a higher level, such as by training a fusion model that takes the feature vectors as input and produces a fused representation. The fused representation can then be used as input to a downstream task-specific model. This approach allows for more complex interactions between the modalities but requires more data and computational resources.

iii) Multi-task learning: This involves training a single model to perform multiple tasks, where each task is associated with a different modality. This can be useful when the modalities are related and when the shared representations learned by the model can improve performance on all tasks.

iv) Attention mechanism: Attention mechanisms allow the model to selectively focus on certain parts of the input data, which can be particularly useful when the different modalities contain complementary information. Attention mechanisms can be used as a form of feature fusion, where the attention weights learned by the model indicate the importance of each modality for a given task.

v) Generative models: Some multimodal tasks such as image captioning, text-to-speech, etc. can benefit from using Generative models such as GANs, VAEs etc. to generate one modality from another.

Our Process Model is as follows: Live data i.e., Audio, Video & Text data of an individual is collected and passed through the respective pre-trained model. Finally, 8 basic emotions according to Platchik model is predicted.

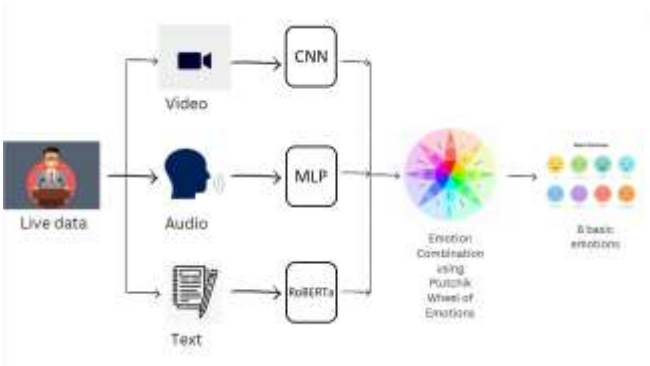

Fig. 2: Our Process Flow Model

Figure3: Pipeline Diagram of the various stages of Emotion Detection in our Model

Input Data Block	Transfer Learning Block	Combining the output predicted emotion of individual TL model to Final Classifier(SVM/ DT/Ensemble)	Final Emotion Detection Block
Stage 1	Stage2	Stage 3	Stage 4

Phase 1: Step1: Collect Multimodal Data Collection phase:

a) Video data from camera of the device (Laptop, Mobile….)

b) Voice Data from Microphone of the device

c) Text Data from the Manuscripts or Blogs (By web scarping/content scraping data using Tools like UIPath or Blue Prism)

d) Many Physiological Data such as Biomarkers data ECG / Skin conductance, Heart Rate, Movement, Speech Patterns

Input Data $_{video}$
Input Data $_{acoustic}$
Input Data $_{text}$
Input Data $_{ecg}$
.....

…….

(ID video, ID accostic, ID text, ID ecg)

Phase 2/Step 2: Usage of Transfer Learning (Pre Trained Models)

a) Transfer Learning (For Video take Pretrained Models like CNN / VGG16/ VGG32 available in Keras/ Tensor flow Library

b) For Text data Use pertained Models like BERT / Distil BERT / RoBERT/ AlBERT available in Hugging Face Library

c) For Voice Take Pretrained Model such as DEAP, E-LSTM,

Transfer Learning $_{\text{video}}$
Transfer Learning acoustic
Transfer Learning text
Transfer Learning ecg

.....

…….

(TLvideo,TLAccostic, TL text, TL ecg)

Step3:

Step3: Combining the Individual TL output to classifier like SVM/Decision Tree/ Random Forest/ Ensemble (Bagging / Boosting) to Final result by Multiplicative/ additive or Concatenation / methods

Phase 4/Step4 Final Prediction of Emotions: Combining the Individual TL output to classifier like SVM/Decision Tree/ Random Forest/ Ensemble (Bagging/ Boosting) to Final result by

Multiplicative/ additive or Concatenation / methods (Joint, coordinated or Encoder Decoder Methods)

E_{1joy}
E_{2trust}
$E_{3Surprise}$
E_{4Anger}
$E_{5sadness}$
$E_{6Disgust}$
$E_{7Anticipation}$
E_{8Fear}

Collecting input from various sources of data such as Image, Voice, Physiological data (such as ECG,), Pass it through Individual pretrained Neural network (CNN or and do fine Tuning of parameters at the Output stage of Data Through the Network.)

V. Conclusion

In this work we have considered one of the BERT refined Model called Distill BERT. Distill BERT is smaller, faster, cheaper and lighter version of BERT for text, for video and image popular pre - trained models such as CNN with VGG16 have been used. For acoustic data we are using Librosa Python Library and finally not early but late combination of various model is being done. Final different type of emotions detected is based on Platchik's model ie "wheel of Emotion Model". Future scope of the work can be towards transformer/ Attention mechanism/ (Product Comparison) ie finding Common vector space & Reducing Heterogeneity Gap

References

[1] Liu, Kuan and Li, Yanen and Xu, Ning and Natarajan, Prem ,"Learn to Combine Modalities in Multimodal Deep Learning" 2018, https://doi.org/10.48550/arxiv.1805.11730.

[2] S. Sah, A. Shringi, D. Peri, J. Hamilton, A. Savakis and R. Ptucha, "Multimodal Reconstruction Using Vector Representation," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, pp. 3763-3767, 2018, doi: 10.1109/ICIP.2018. 8451439.

[3] Wenzhong Guo, Jianwen Wang, and Shiping Wang, "Deep Multimodal Representation Learning: A Survey", date of publication May 15, 2019, date of current version May 28, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2916887.

[4] Thanyathorn Thaanapattheerakul, Jacqueline Amoranto, Jonathan H. Chan, "Emotion in a Century: A Review of Emotion Recognition", IAIT 2018: Proceedings of the 10th International Conference on Advances in Information Technology December 2018 Article No.: 17Pages 1–8https://doi.org/10.1145/3291280.3291788

[5] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Ukasz and Polosukhin, Illia, "Attention is All you Need", {https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c43f5ee243547dee91fbd053c1c4a845aa-Paper.pdf, volume 30, year 2017, https://doi.org/10.48550/arXiv.1706.03762

[6] Liu Dong, Wang Zhiyong, Wang Lifeng, and Chen Longxi, "Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning," Frontiers in Neurorobotic, vol. 15, 2021, URL=https://www.frontiersin.org/articles/10.3389/fnbot.2021.697634 ,OI=10.3389/fnbot.2021.697634 ,ISSN=1662-5218

[7] W. Lin, I. Orton, Q. Li, G. Pavarini and M. Mahmoud, "Looking At The Body: Automatic Analysis of Body Gestures and Self-Adaptors in Psychological Distress," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2021.3101698.

[8] S. Koelstra et al., "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," in IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18-31, Jan.-March 2012, doi: 10.1109/T-AFFC.2011.15.

[9] Wei Yao, Anastasia Moumtzidou, Corneliu Octavian Dumitru, Stelios Andreadis, Ilias Gialampoukidis, Stefanos Vrochidis, Mihai Datcu, and Ioannis Kompatsiaris, "Early and Late Fusion of Multiple Modalities in Sentinel Imagery and Social Media," Retrieval. In: 2021, Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science(), vol 12667. Springer, Cham. https://doi.org/10.1007/978-3-030-68787-8_43

[10] Chul Min Lee, S. S. Narayanan and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs," Proceedings. IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, vol. 1, pp. 737-740 2002, doi: 10.1109/ ICME.2002.1035887.

[11] M. A. Ullah, M. M. Islam, N. Binti Azman and Z. M. Zaki, "An overview of Multimodal Sentiment Analysis research: Opportunities and Difficulties," 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Dhaka, Bangladesh, pp. 1-6, 2017, doi: 10.1109/ICIVPR.2017.7. 890858.

[12] Han, Kun, Yu, Dong and Tashev, Ivan, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine", Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 10.21437/ Interspeech.2014-57.

[13] Sitharthan, R., Vimal, S., Verma, A., Karthikeyan, M., Dhanabalan, S. S., Prabaharan, N., ...&Eswaran, T. (2023). Smart microgrid with the internet of things for adequate energy management and analysis.Computers and Electrical Engineering, 106, 108556.

[14] Zhaoxia Wang, Seng-Beng Ho, Erik Cambria, "A review of emotion sensing: categorization models and algorithms", Springer Science ,Business Media, LLC, part of Springer Nature, 2019.

[15] Shaver, P. Shaver, J. Schwartz, D. Kirson, O'Connor C, "Emotion knowledge: further exploration of a prototype approach," J Pers Soc Psychol, Parrott WG, vol. 52, no. 6, pp.1061–1086 6, 2001.

[16] C. Adam, A. Herzig, and D. Longin, "A logical formalization of the OCC theory of emotions," Synthese, vol. 168, pp. 201–248, 2009,. https://doi.org/10.1007/s11229-009-9460-9.

[17] Paul Ekman, "An argument for basic emotions Cognition and Emotion," vol. 6, no. 3-4, pp. 169-2000, Published online: 07 Jan 2008, https://doi.org/10.1080/02699939208411068.

[18] Z. Wang, S.B. Ho, and E. Cambria, "A review of emotion sensing: categorization models and algorithms," Multimed Tools Appl., vol. 79, pp. 35553–35582, 2020, https://doi.org/10.1007/s11042-019-08328-z

[19] Neviarouskaya, Alena, Prendinger, Helmut and Ishizuka, Mitsuru, M.: Textual Affect Sensing for Sociable and Expressive Online Communication. Lecture Notes in Computer Science, vol.. 4738, pp. 218-229, 2007, 10.1007/978-3-540-74889-2_20.

[20] W. Han, T. Jiang, Y. Li, B. Schuller and H. Ruan, "Ordinal Learning for Emotion Recognition in Customer Service Calls," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6494-6498, doi: 10.1109/ICASSP40776.2020.9053648.

[21] P. Lakhan et al., "Consumer Grade Brain Sensing for Emotion Recognition," in IEEE Sensors Journal, vol. 19, no. 21, pp. 9896-9907, 1 Nov.1, 2019, doi: 10.1109/JSEN.2019.2928781.

[22] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A, Zadeh, and L.P. Morency, "Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors," Proc Conf AAAI Artif Intell. Vol. 33, no. 1, pp. 7216-7223, Jul 2019, PMID: 32219010; PMCID: PMC7098710.

[23] Tadas Baltrusaitis, Chaitnya Ahuja, and Louis Philippe Morency, "Multimodal Machine Learning: A Survey and Taxonomy",IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41 issue 2, pp. 423–443, February 2019 https://doi.org/10.1109/TPAMI.2018.2798607.

[24] Siriwardhana, Shamane, Kaluarachchi, Tharindu, Billinghurst, Mark, Nanayakkara, and Suranga, "Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion", IEEE Access, vol. 8, pp. 176274-176285, January 2020, DOI:10.1109/ACCESS.2020.3026823.

[25] Liu, Kuan and Li, Yanen and Xu, Ning and Natarajan, Prem "Learn to Combine Modalities in Multimodal Deep Learning", 2018, https://doi.org/10.48550/arXiv.1805.11730.

[26] Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron, Salakhutdinov, Ruslan, Zemel, Richard and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning – vol. 37, pp. 2048–2057, 06 July 2015.

[27] K.A. Araño, Gloor, Orsenigo, "When Old Meets New: Emotion Recognition from Speech Signals" Cogn Comput., vol. 13, pp.771–783, 2021, https://doi.org/10.1007/s12559-021-09865-2.

[28] Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0.Computers and Electrical Engineering, 104, 108481.

[29] Plutchik, Robert. "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice." American scientist 89.4, pp. 344-350, 2001.