Evaluating Significant Features in Context-Aware Multimodal Emotion Recognition with XAI Methods

Talal Shaikh[1], Aaishwarya Khalane[1], Rikesh Makwana[1], and Abrar Ullah[1]

[1]Heriot-Watt University - Dubai Campus

January 18, 2023

Abstract

Analysis of human emotions from multimodal data for making critical decisions is an emerging area of research. The evolution of deep learning algorithms has improved the potential for extracting value from multimodal data. However, these algorithms do not often explain how certain outputs from the data are produced. This study focuses on the risks of using black-box deep learning models for critical tasks, such as emotion recognition, and describes how human understandable interpretations of these models are extremely important. This study utilizes one of the largest multimodal datasets available - CMU-MOSEI. Many researchers have used the pre-extracted features provided by the CMU Multimodal SDK with black-box deep learning models making it difficult to interpret the contribution of individual features. This study describes the implications of individual features from various modalities (audio, video, text) in Context-Aware Multimodal Emotion Recognition. It describes the process of curating reduced feature models by using the GradientSHAP XAI method. These reduced models with highly contributing features achieve comparable and even better results compared to their corresponding all feature models as well as the baseline model GraphMFN proving that carefully selecting significant features can help improve the model robustness and performance and in turn make it trustworthy.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Evaluating Significant Features in Context-Aware Multimodal Emotion Recognition with XAI Methods

Aaishwarya Khalane | Rikesh Makwana | Talal Shaikh* | Abrar Ullah

[1]School of Mathematical and Computer Science, Heriot-Watt University, Dubai, UAE

Correspondence
*Talal Shaikh, School of Mathematical and Computer Sciences, Heriot-Watt University, Dubai Knowledge Park, Dubai, UAE. Email: t.a.g.shaikh@hw.ac.uk

Summary

Analysis of human emotions from multimodal data for making critical decisions is an emerging area of research. The evolution of deep learning algorithms has improved the potential for extracting value from multimodal data. However, these algorithms do not often explain how certain outputs from the data are produced.

This study focuses on the risks of using black-box deep learning models for critical tasks, such as emotion recognition, and describes how human understandable interpretations of these models are extremely important. This study utilizes one of the largest multimodal datasets available - CMU-MOSEI. Many researchers have used the pre-extracted features provided by the CMU Multimodal SDK with black-box deep learning models making it difficult to interpret the contribution of individual features. This study describes the implications of individual features from various modalities (audio, video, text) in Context-Aware Multimodal Emotion Recognition. It describes the process of curating reduced feature models by using the GradientSHAP XAI method. These reduced models with highly contributing features achieve comparable and even better results compared to their corresponding all feature models as well as the baseline model GraphMFN proving that carefully selecting significant features can help improve the model robustness and performance and in turn make it trustworthy.

KEYWORDS:
Multimodal Emotion Recognition, Explainable Artificial Intelligence, Interpretability, XAI, CMU-MOSEI, BERT

1 | INTRODUCTION

In the recent years, Deep neural network (DNN) has emerged as an important machine learning tool to accomplish high performance on many learning tasks comparable to humans. However, deep learning models are inherently black-box and outputs are often produced with no interpretation or explanation to understand the aspects in the input that influenced the decisions of the model. These systems and decisions can be found in high risk and critical domains such as health, law and order, automotive etc. Given the nature of decisions, it is important for humans to understand the dominant features contributing to the DNN output in a specific context.

Human emotion recognition is an important and ongoing research area. In human emotion recognition application scenarios, various deep and shallow models interpret human emotions to provide various services like controlling appliances. In literature, extensive research has been done on human emotion recognition [1,2]. In these research works, one of the largest datasets evaluated is the CMU MOSEI Dataset [3]. This massive dataset contains many real world, un-staged videos depicting human emotions in a multimodal format comprising of audio, video and text modalities. We observed that many studies [4,5,6,7,8,9] use the same pre-extracted features provided by the CMU-MOSEI Multimodal SDK toolkit. However, it is

essential that AI be transparent about the reasoning used in order to increase trust, clarity, and understanding of these applications. This study aims to determine which features influence the prediction capabilities of the model. Subsequently, we attempt to evaluate the effect of reducing the features to a subset consisting of highly contributing features on the performance of the models.

2 | BACKGROUND RESEARCH

2.1 | Explainable Artificial Intelligence (XAI)

The success of AI in delivering robust solutions has led to its extensive use in applications such as Emotion Recognition (ER), smart ecosystems, smart learning, finance, security, etc. This can be attributed to the ability of AI to enable improved productivity, better decision making, reduction of expenditures, and improved risk management. However, the techniques used for developing these AI solutions such as deep learning often do not explain how or why the certain output is obtained. These Blackbox/opaque models with large amounts of high-dimensional feature vectors output a final result without any human intelligible interpretation of the internal logic applied in these processes [10]. Lacking such auditability in AI systems can prove to be ethically risky and hazardous in real-world applications impacting the safety of the users [11].

To develop and deploy trustworthy AI solutions, carefully balancing the trade-off between the prediction and explainability of these systems is essential. As seen in Figure 1, the explainability of machine learning models is inversely proportional to their prediction accuracies [12]. While deep learning models are known for their ability to achieve high prediction accuracies with minimal need for human intervention, they are accompanied by the curse of being highly opaque and un-interpretable by humans. Explainable AI (XAI) systems help tackle this by helping the user understand the behind-the-scenes processing logic of deep learning AI systems using simple, interpretable models [13]. XAI methods help decode these inexplicable, uninterpretable black boxes into transparent, human interpretable glass boxes.

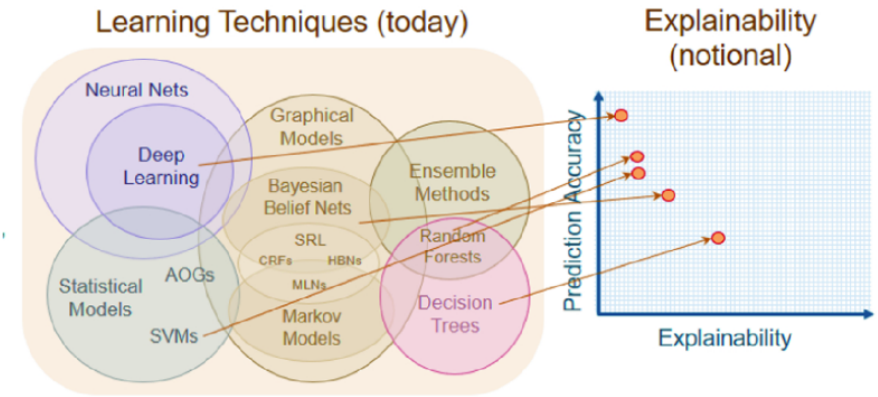

Figure 1 Explainability v/s Prediction Accuracy of ML Models [12]

Interpretable machine learning techniques are grouped into two categories: intrinsic and post-hoc methods [14]. Intrinsic interpretability: Using simple interpretable models which are self-explanatory from their internal structure. Examples of such include decision trees, linear models, etc [14]. Post-hoc interpretability: Complex black-box models can be interpreted after model training (post hoc) using a model-agnostic surrogate model [14]. Model-agnostic methods work by changing the input of the machine learning model and measuring changes in the prediction output [14,15]. The surrogate can either be global or local. The global surrogate model approximates the overall prediction of the black-box model, whereas the Local surrogate model explains individual predictions of the black-box model [14,15]. Figure 2 provides an overview on the above discussed methods.

2.2 | Emotion Recognition Overview

Understanding and responding to emotions is an integral part of human communication. Emotions thus play a crucial role in Human-Computer Interaction (HCI). Therefore, extensive research has been conducted to develop intelligent systems capable of recognising and understanding human emotions as organically and efficiently as possible using Emotion Recognition (ER) algorithms. ER finds its applications in both simple day-today systems including smart mirrors, customer satisfaction, gaming, chat-bots, smart home solutions as well as more complex, critical systems such as healthcare, criminal activity detection, mental health monitoring, emotion recognition of drivers for maintaining road safety, etc. Such

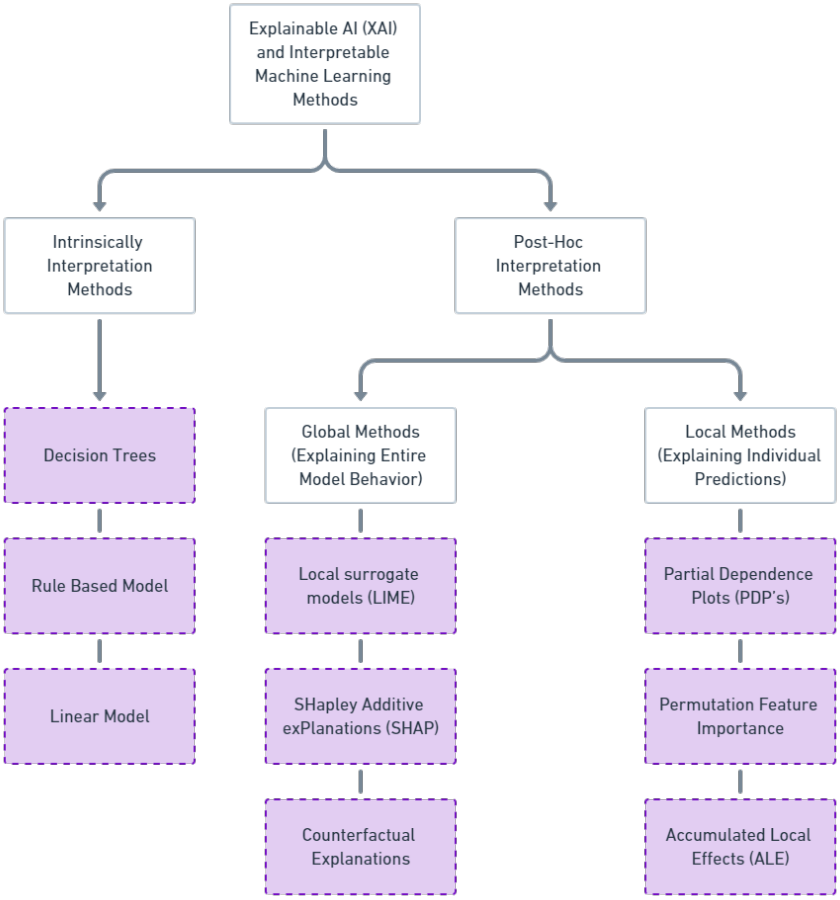

Figure 2 Common Explainable AI (XAI) and interpretable Machine Learning (ML) techniques. Adopted from [14,16]

critical applications of ER can be extremely sensitive to the final results obtained from the ER process. Incorrectly detecting emotions in such scenarios can be extremely hazardous and cause serious repercussions. Hence meticulously modelling the ER process and making the black-box machine learning models explainable to and interpretable by humans is crucial.

2.3 | Unimodal v/s Multimodal Emotion Recognition

Recognizing emotions is not a straightforward task. Emotions are naturally perceived by humans using a fusion of various cues like facial expressions (visual), voice modulation (acoustic), words spoken(Textual). Unimodal ER techniques can prove insubstantial [1] in scenarios like sarcasm - the sarcastic expression of a disappointed smiling face could be classified as "happy" if the only focus is the visual cue. We hence focus exclusively on using multiple modalities (bimodal and trimodal ER) as opposed to single modalities (unimodal ER) to infer the appropriate emotion from a video as the multimodal ER technique reflects nuances of real emotional perception and makes the ER system more robust and reliable [18].

2.4 | CMU MOSEI Dataset

CMU MOSEI (Multimodal Opinion Sentiment and Emotion Intensity)[3] is the largest emotion recognition and sentiment analysis dataset available with 23,453 video segments from 3228 distinct videos featuring 1000 subjects talking about 250 diverse topics. This open-source dataset is comprised of natural (un-staged) videos procured from platforms like Youtube and have been quality-checked 14 judges across all three modalities (audio,visual and textual). Our analysis uses the pre-extracted features available in the CMU-Multimodal SDK [19]. The visual features were extracted at 30Hz, 35 features per timestep. After extracting facial action units and several facial landmarks, the final visual embeddings were extracted by utilizing deep learning frameworks such as FaceNet, Deep-Face, and SphereFace. Using static facial images, Emotient FACET was employed to extract the six emotions (happy, sad, angry, surprise, disgust, fear). Covarep was employed to extract 74 CMU MOSEI audio features per timestep

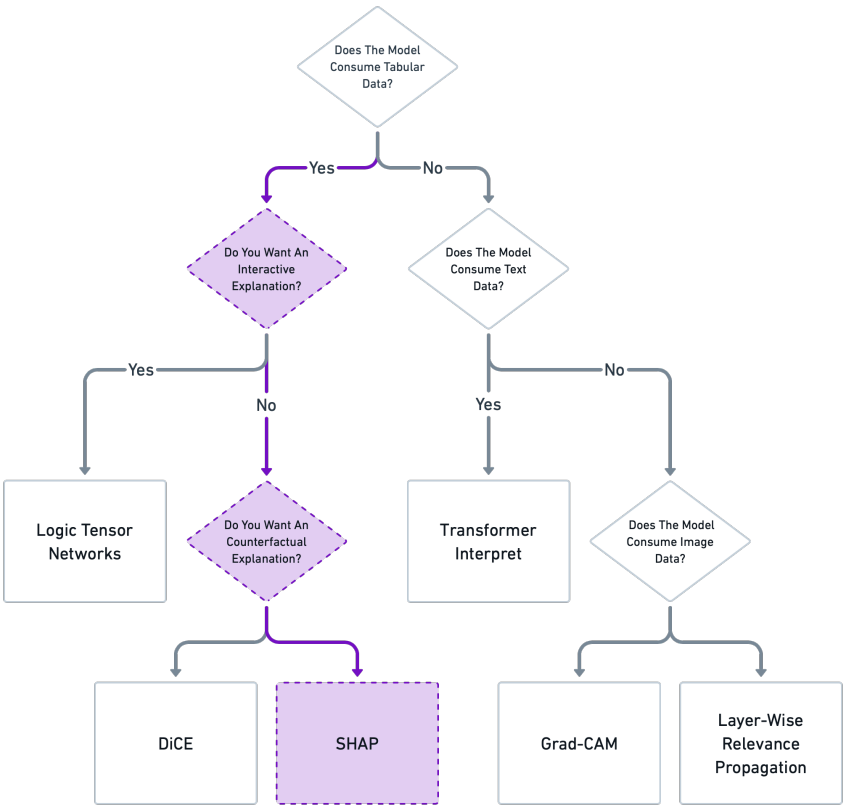

Figure 3 Flowchart to determine which explainable methods to use, adopted from [17]. The highlighted path describes the reason for selecting SHAP as our method for this study.

such as 12 MFCCs, pitch, maxima dispersion quotients, etc. to portray emotions described by speech tonality. 300 textual features were extracted using GloVe embeddings.

In our research on the CMU-MOSEI dataset, we came across multiple papers [4,5,6,7,8,9] using the pre-extracted features provided by the CMU-Multimodal SDK [19] . Most of these papers used the same 74 audio and 35 visual features as the baseline paper [3].

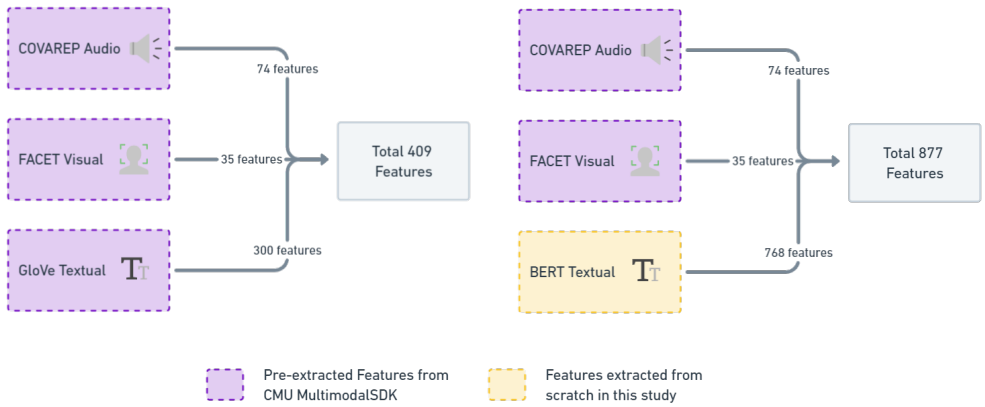

Figure 4 Feature Splits used in this study include pre-extracted features acquired from CMU MultimnodalSDK [19] and features extracted from scratch inspired by [20]

As for the textual modality, we came across a multitude of papers shifting to using BERT for extracting textual features rather than using the provided GloVe embeddings [21,22,23,24]. BERT or Bidirectional Encoder Representations from Transformers is a transformer-based model widely used for extracting high-quality textual embeddings. Conventional word embeddings like GloVe construct a single word vector for each unique word whereas BERT uses a bidirectional attention mechanism to recognize contextual information. Moreover, the pre-trained BERT model proves to be convenient with its pre-encoded language information which facilitates quicker development using high-quality features even with the availability of smaller training data.

2.5 | Problem with pre-extracted features

The CMU MOSEI dataset is one of the most popular datasets used for multimodal emotion recognition. It has been heavily referenced with 129 citations on Scopus from which 35 papers [25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59] directly utilize CMU MOSEI features in their application, most of which utilize deep-learning architectures for their analysis. This existing research on the CMU MOSEI dataset however does not explore the explainability of the CMU MOSEI features. The pre-extracted features provided by the CMU MOSEI SDK have not been named/described. For applications directly using CMU features in their application or devices, it is highly important to understand/interpret these anonymous features - how they were generated, what they truly symbolize, or how they contribute to the final results. There has been prior research to make the pre-extracted features interpretable by using shallow-learning methods [3,60]. However, these features when used with deep-learning models, are essentially black-box features that we have no information about and hence cannot be used to comprehend the behavior of the models. We, therefore, felt the need to devise a way to understand the behavior of these features and the impact of their attributions in our deep-learning model. Using an XAI model to explain and interpret the attribution of features of each modality can help to improve the understanding of the feature significance and help us decide which modality contributes the most to ER. This can ultimately lead to improved robustness and performance accuracy in detecting the appropriate emotion and consequently minimize the hazards involved in wrongly identifying emotions in many scenarios such as criminal investigations, mental health monitoring, etc.

3 | METHODOLOGY

Our experiments used deep-learning to recognize emotions using the audio, visual and textual modalities by utilizing the pre-extracted audio-visual features by combining them with the pre-extracted GloVe as well as newly extracted BERT features. We used an early fusion mechanism to fuse the three modalities which were then passed through a Bi-LSTM for emotion classification. We used a two-layered bidirectional LSTM having a hidden layer with 256 neurons inspired by [20] (Figure 5).

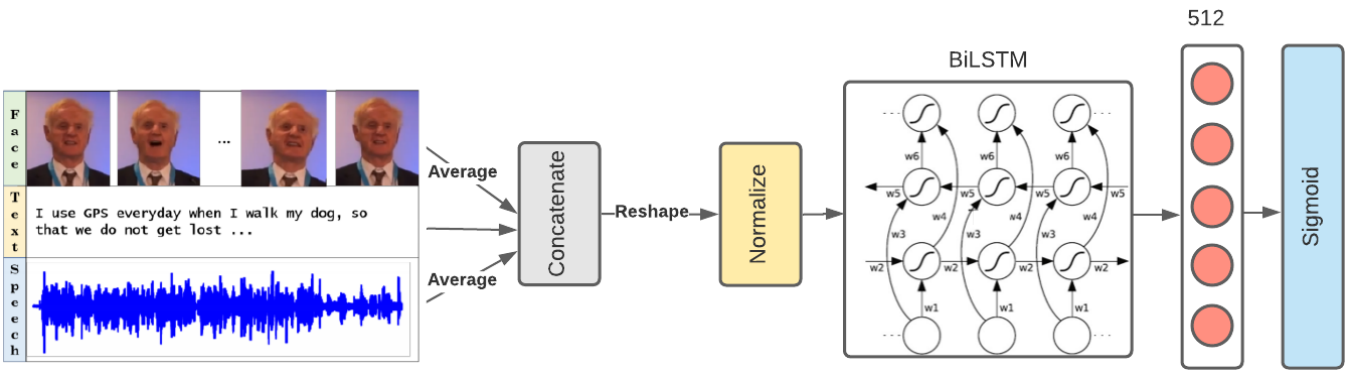

Figure 5 Our Early Fusion with BiLSTM model [20]

The first set of experiments involved training the Bi-LSTM model with early-fused 409 trimodal features including the pre-extracted 74 audio (A1-A74) and 35 visual (V1-V35) and 300 textual GloVe features (T1-T300) from the CMU Multimodal SDK [19]. Taking inspiration from [21,22,23,24], we then extracted 768 BERT features per sentence by fine-tuning the pre-trained Base BERT model (provided by the Huggingface Transformers library). We then fused these 768 BERT textual features (T1-T768) with the pre-extracted 74 audio (A1-A74) and 35 visual (V1-V35) features and trained new Bi-LSTM models using these 877 features.

Our ultimate aim was to interpret our models using XAI at the feature-level to find the importance of each individual feature with respect to their contribution to the model output also known as Primary Attribution. We needed to choose a suitable XAI method for our primary attribution analysis from the vast array of existing XAI methods.

[17] suggest the use of suitable explainable AI methods depending on the type of data used as illustrated in Figure 3. The data consumed by our model is tabular. While counterfactual explanations for a black-box model help understand how the smallest change to the feature values changes the prediction to a predefined output [14], we needed to understand the overall global feature contribution for the model. We hence decided to use SHAP or SHapley Additive exPlanations method to interpret our model.

SHAP (SHapley Additive exPlanations) is a post-hoc model-agnostic method that explains individual predictions by assigning each feature an importance value for a particular prediction [61 62]. The Shapley value attribution method is inspired by a cooperative game theory concept. It takes each permutation of the input features and individually adds them to the provided baseline. In this process, the output difference after adding each of the features corresponds to the contribution of that feature, and the average of these differences across all permutations determines the attribution. Due to the multiple permutations involved, using a larger number of features makes this method computationally intensive [63].

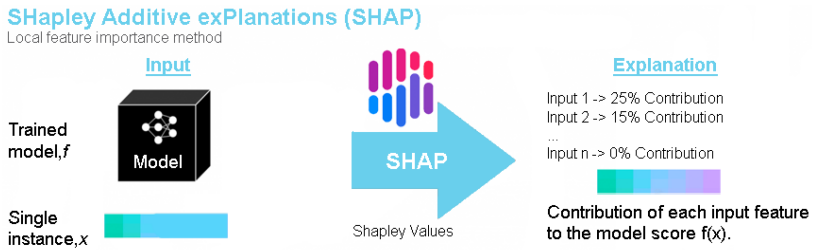

Figure 6 SHAP (SHapley Additive exPlanations) pipline for explaining single predictions [62]

To interpret the importance of the attribution of these features to the final classification results, we tested our models with GradientSHAP method provided by the Captum library.

3.1 | Captum and GradientSHAP

Captum [63] is an open-source XAI library for PyTorch which helps with the explainability and interpretability of various AI models. Captum provides algorithms for evaluating attribute (feature, neuron and layer) importance and support for multimodality inputs such as text, audio and video [63]. This study focuses on identifying the significant effects of input modality by examining the attribution value (relevance or contribution) of the input features to the deep neural network. Captum provides two categories of attribution methods, perturbation and gradient-based.

1. Perturbation-based Methods: Compute the attribution value for an individual or a set of input features by perturbing (removing, masking or altering) them before performing a forward pass [64]. Finally, calculate the difference between the new and original output [64].

2. Gradient-based Methods: Compute the attribution values for all input features on a single forward and backward pass [64]. However, unlike the perturbation-based methods, attributes can not always be directly related to changes in output.

GradientSHAP available in Captum is a gradient algorithm used to compute SHAP values to evaluate the primary attribution of models. In gradient SHAP, each input sample is subjected to multiple Gaussian noise additions, a random point is selected along the path between baseline and input, and the gradient is computed based on the chosen random points. The algorithm assumes that the input features are independent and the explaining model is linear between the provided baselines and input features. It results into attributions approximating SHAP values that denote the expected value of **gradients * (inputs - baselines)** [63].

We employed GradientSHAP and experimented with various subsets (top5, top10, top15, top20, top25) of the features with highest absolute attribution values. From these subsets, we empirically established that the optimal subset of features is top 20 features to obtain performance results close to all feature models while being inclusive of all modalities to find the most highly contributing features. We hence found out the top 20 features with the greatest absolute attribute values (check figure 7 and section 4.1 for reference). To gauge and validate the contribution of these top 20 features towards the classification results, we used these 20 features to train new models. We then tested these newly trained reduced feature models and compared their performance with the models trained with all features (Section 4.2).

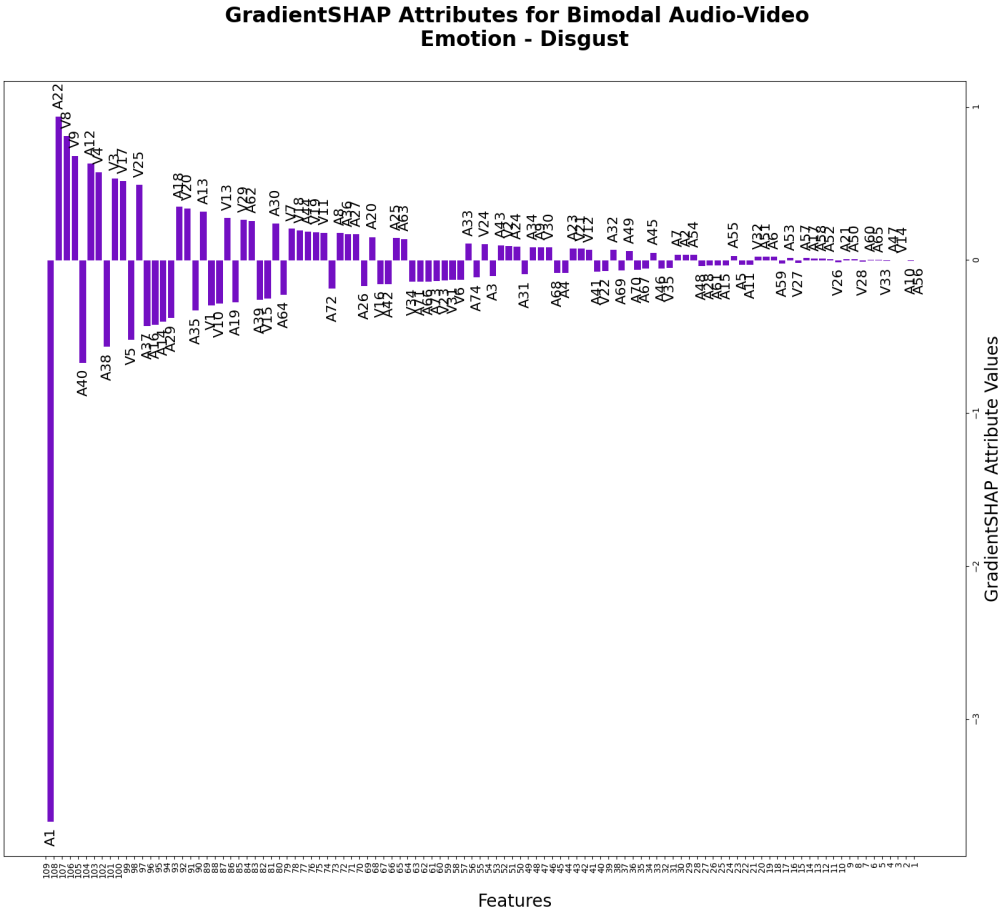

Figure 7 GradientSHAP Attribute Values for the features of the emotion Disgust from the Bimodal (A,V) model sorted from highest to lowest (refer table 1)

4 | RESULTS

The results obtained from these experiments were as follows. The modalities have been referred to with their initials to keep the document concise - A for Audio, V for Visual and T for Textual.

4.1 | Top 20 Features based on GradientSHAP Attribute Importance

The tables 1, 2, 3, 4, 5, 6, 7 list the top twenty features with the highest absolute attribution scores along with the percentage contribution pf the respective modalities involved as well as the modality which dominates for 7 different bimodal and trimodal models. As observed from these findings, the dominant modality was found to be Text for the majority of the experiments followed by Video (Visual) and Audio which was found to be the least dominating modality when considering the top 20 features. Please check the **??** for visual illustration of the contribution of various modalities in terms of Top 20 Features.

4.2 | Discussing the Results obtained by All Feature Models v/s Top 20 Feature Models

We evaluate the models based on the metrics used by our baseline paper[3] - F1 Score and Weighted Accuracy. The Weighted Accuracy metric (figure 8) is used to avoid any discrepancies caused due to the imbalance in the CMU-MOSEI dataset[3].

Table 1 Bimodal (Audio, Video) Analysis

Emotion	Top 20 Features	Audio	Video	Dominant Modality
Fear	A1, V34, V5, V9, V7, A65, A12, A63, V14, V13, V17, V1, V10, A67, A13, A9, A66, V23, A64, V33	45%	55%	VIDEO
Sad	A1, A12, V7, V9, V6, V3, V14, V1, V4, V29, V22, V24, V13, A20, V23, A19, V2, V34, A39, A28	30%	70%	VIDEO
Angry	A1, A12, V9, V1, V8, V10, V23, V6, A13, V2, V4, V17, V29, V7, A14, A19, V33, V25, A18, V5	30%	70%	VIDEO
Disgust	A1, A22, V8, V9, A40, A12, V4, A38, V3, V17, V5, V25, A37, A16, A14, A29, A18, V20, A35, A13	60%	40%	AUDIO
Happy	A1, A12, V9, V4, V6, V7, V19, V10, V1, V15, V14, V23, V5, V26, V20, V16, V8, V17, A13, V30	15%	85%	VIDEO
Surprise	A12, A1, V4, V8, A20, V3, A13, A39, A29, A42, A27, A28, V21, V14, A43, V17, A22, V19, A30, V2	60%	40%	AUDIO

$$\text{WAcc.} = \frac{TP \times N/P + TN}{2N}$$

Figure 8 Weighted Accuracy Formula where N is the total number of negative labels while P is the total number of positive labels. TP represents True Positives while TN represents True Negatives.

4.2.1 | F1 Scores

We compare the F1 scores obtained by reduced feature models with Top 20 Features (Table 9) against the models trained with with All Features (Table 8). The F1 score results for the Bimodal(A,V) analysis show that the Top 20 feature model outperformed the All Feature model in 3 emotions (Disgust, Happy and Surprise) while being very close for the remaining three emotions (Sad, Angry and Fear).

Comparing all versus top 20 features for GloVe models, we observe that for Bimodal(T,A) analysis, we achieved the same F1 scores for 4 out of 6 emotions, except Sad and Happy, for which results obtained by using top 20 features were still very close to those obtained by using all features. For Bimodal(T,V), the F1 score performance of top 20 feature model was the same for 2 (Fear, Suprise) out of six emotions, slightly worse for 3 (Sad, Angry Disgust) and slightly better for the Happy class as compared to the All Features model. For the Trimodal analysis, the All Feature model outperformed the Top Twenty features model in 4 (Sad, Angry, Disgust, Happy) out of 6 emotion classes while obtaining exactly same F1 scores for Fear and Disgust.

For BERT models, we observe that for Bimodal(T,A), the Top 20 Features model performed slightly better than the All Features model in 3 (Angry, Disgust, Happy) out of 6 emotion classes. The F1 scores of Top 20 Features were still very close to those obtained by the all feature model for the remaining three emotions (surprise, fear, sad). For Bimodal(T,V), the F1 score performance of top 20 feature model was the same for fear and slightly better for angry and disgust but slightly worse for the remaining 3 emotions (sad, happy, surprise) as compared to the all feature model. For the Trimodal analysis both the models had the same F1 score for fear and surprise, the top twenty model performed better for angry and worse for sad, disgust and happy classes.

Our reduced feature models with Top 20 Features were able to obtain better F1 scores than the baseline GraphMFN model (using all 409 features)[3] for 4 out of 6 emotions (Sad, Angry, Disgust and Happy). For the emotion class Sad, our BERT Bimodal (T,A), BERT Trimodal and Bimodal (A,V) outperformed the baseline. Three of our Top 20 features models BERT Bimodal(T,A), BERT Bimodal(T,V) and BERT Trimodal outperformed the baseline for the emotions Angry and Disgust. For the emotion Happy, four of our models including BERT Bimodal(T,A), BERT Bimodal(T,V), BERT Trimodal and Bimodal(A,V) outperformed the baseline. Surprisingly, all six of our Top 20 Features models obtained better F1 scores than the baseline GraphMFN[3].

Table 2 GloVe Bimodal (Text, Video) Analysis

Emotion	Top 20 Features	Text	Video	Dominant Modality
Fear	T158, T225, T101, V17, T25, T146, T84, V3, T90, T132, V7, T62, V21, T156, T200, T261, T293, T229, T210, T250	80%	20%	TEXT
Sad	V3, V9, V7, T28, T44, T156, V4, T251, V14, T233, T160, T267, T214, T192, T123, V17, T212, T117, T84, V24	65%	35%	TEXT
Angry	T71, T36, T271, T143, T232, T293, T170, T282, T241, V21, T98, T246, T214, T34, T105, T49, T260, T136, T119, V1	90%	10%	TEXT
Disgust	T232, T206, T271, T160, T143, T170, T214, T71, T256, T8, T28, T156, T101, T197, T36, T261, T223, T158, T240, T260	100%	0%	TEXT
Happy	T82, T143, V4, T105, T251, V7, T152, T271, V2, T267, T12, T36, T185, T28, T96, T49, T15, T230, T232, T240	85%	15%	TEXT
Surprise	T202, T167, T282, T249, T182, T293, T16, T287, T121, T191, V7, T220, T156, T20, T242, T265, T79, T70, T14, T274	95%	5%	TEXT

Table 3 BERT Bimodal (Text, Video) Analysis

Emotion	Top 20 Features	Text	Video	Dominant Modality
Fear	V34, V1, V23, T763, V8, V7, T675, V5, T27, T720, T149, V4, V33, T606, V9, T370, T68, T564, T335, T468	55%	45%	TEXT
Sad	V7, V4, V9, V8, V23, V6, V14, V24, V5, V13, T323, V2, V34, V19, V17, V3, V29, T723, V10, T288	15%	85%	VIDEO
Angry	V4, V1, V8, V3, V7, V21, V9, T79, T713, V15, T675, T244, T623, V10, V24, T763, V19, T149, T70, T142	45%	55%	VIDEO
Disgust	V17, V34, V8, V29, V23, V1, V2, V7, V21, T675, T135, T125, T133, T623, V5, T96, T699, T571, T53, T90	50%	50%	TEXT , VIDEO
Happy	V4, V7, V9, V1, V6, V19, V10, V15, V23, V25, V16, V20, T44, T763, V8, V34, V24, T26, T203, T507	25%	75%	VIDEO
Surprise	V8, V9, V14, V2, T181, V5, T271, V12, T12, V28, T763, T402, T121, T708, T234, T711, T41, T623, T573, V10	60%	40%	TEXT

4.2.2 | Weighted Accuracies

We compare the weighted accuracies obtained by reduced feature models with Top 20 Features (Table 11) against the models trained with with All Features (Table 10). The Weighted Accuracies obtained in the Bimodal(A,V) analysis show that the Top 20 feature model outperformed the All Feature model in 2 emotions (Disgust and Happy), tied for the emotions Fear and Surprise while being very close for the remaining two emotions Sad and Angry.

Comparing all versus top 20 features for GloVe models, we observe that for Bimodal(T,A) analysis, we achieved the same weighted accuracies for 4 out of 6 emotions, except Sad and Happy, for which results obtained by using top 20 features performed slightly worse (but still very close) compared to the all feature model. For Bimodal(T,V), the weighted accuracies of top 20 feature model was the same for 2 (Fear and Suprise) out of six emotions, slightly worse for 3 (Sad, Angry, Disgust) and slightly better for the happy class as compared to the all feature model. For the Trimodal analysis, the all feature model outperformed top twenty model feature model in 4 (Sad, Angry, Happy and Surprise) out of 6 emotion classes while giving exactly same F1 scores for Fear and Disgust. The trends with weighted accuracies are exactly the same (identical) as observed with the F1 score comparison between the two Bimodal(T,A), Bimodal(T,V) and trimodal models.

Table 4 GloVe Bimodal (Text, Audio) Analysis

Emotion	Top 20 Features	Text	Audio	Dominant Modality
Fear	A1, A12, T158, T146, A31, T250, A26, T161, T90, T172, A32, T93, A39, T84, T101, A30, T132, T295, T225, A43	60%	40%	TEXT
Sad	A1, A12, T73, T156, T280, T160, A39, T192, T104, T233, T251, T84, T44, T129, T40, T143, T232, A20, T62, T28	80%	20%	TEXT
Angry	A1, A12, T143, T71, T271, T36, A18, T256, T232, T293, T98, T84, T246, T169, T136, T274, A19, T160, A17, T170	75%	25%	TEXT
Disgust	A1, T232, T143, T160, T271, A22, T214, T170, T101, T254, T28, T256, T71, T240, T1, T206, T197, T239, T36, T106	90%	10%	TEXT
Happy	A1, T271, T280, T143, T256, T214, T129, T104, T258, A18, T105, T246, T101, T89, T267, T251, T68, T36, T185, T146	90%	10%	TEXT
Surprise	A1, A12, T20, T182, T282, T167, T287, A39, T121, T143, T202, T44, T16, A20, A43, T212, T73, A13, T241, T233	70%	30%	TEXT

Table 5 BERT Bimodal (Text, Audio) Analysis

Emotion	Top 20 Features	Text	Audio	Dominant Modality
Fear	A1, A12, T763, T563, T370, T12, T122, T672, T573, T733, T691, T434, T378, T79, T766, T27, T720, T222, T564, T606	90%	10%	TEXT
Sad	A1, A12, A13, A34, A33, A20, A44, A38, T668, A27, A28, A43, A8, A35, T669, A32, T118, A39, A5, A42	15%	85%	AUDIO
Angry	A1, A12, T79, T713, A13, T149, T227, T257, A18, T142, T594, T222, T420, T391, T70, T571, T630, T192, T665, T114	80%	20%	TEXT
Disgust	A1, A12, A22, A17, T161, T560, A33, T121, T94, A40, A32, T573, T763, A28, A34, T135, T99, T402, T36, T151	55%	45%	TEXT
Happy	A1, A12, T44, T162, T377, T206, T413, T26, T763, T507, T203, T742, T200, T353, T80, T27, T622, T766, T234, T327	90%	10%	TEXT
Surprise	A1, T402, T573, T181, T271, T12, T308, A12, T348, T754, T763, T347, T409, T566, T144, T708, T378, T471, T694, T755	90%	10%	TEXT

For BERT models, we observe that for Bimodal(T,A), the top 20 feature model performed slightly better than the all feature model in 3 (Angry, Disgust, Happy) out of 6 emotion classes. The weighted accuracies scores of top 20 features were still very close to those obtained by the all feature model for the remaining three emotions (Surprise, Fear, Sad). For Bimodal(T,V), the weighted accuracy of top 20 feature model was slightly better for Angry and Happy but slightly worse for the remaining 4 emotions (Fear, Sad, Disgust, Surprise) as compared to the all feature model. In the Trimodal analysis, the Top Twenty Features model performed better for Angry and Disgust and slightly worse for Fear, Sad, Happy and Surprise classes. Again, the trends were similar as observed with the F1 score comparison between the Top Twenty Features and All Features models.

Table 6 GloVe Trimodal Analysis

Emotion	Top 20 Features	Audio	Video	Text	Dominant Modality
Fear	A1, A12, V5, V17, V9, T146, V34, V4, T158, A31, V14, V7, T101, T54, T10, T132, T90, A32, T161, T277	20%	35%	45%	TEXT
Sad	V3, V29, A1, V7, V5, V6, T156, A39, V9, T73, T251, T233, T104, T44, V24, T160, T129, T192, T280, A34	15%	35%	50%	TEXT
Angry	A1, A12, T143, V9, V1, V29, T256, T232, T36, T71, T246, T271, V8, T98, A18, T293, T169, V4, T84, T200	15%	25%	60%	TEXT
Disgust	A1, A12, T232, T143, A22, T160, V29, T271, V3, T214, T254, T28, T1, V4, T170, T101, V7, T256, T240, V17	15%	25%	60%	TEXT
Happy	A1, T280, V17, T271, V22, V4, T256, T129, V29, T143, T105, T101, T267, T20, T246, T232, T104, T258, T89, T36	5%	20%	75%	TEXT
Surprise	A1, A12, T167, T20, T182, T282, A39, T202, T143, T287, A42, T121, A43, T16, T241, A20, A27, T50, T28, V14	35%	5%	60%	TEXT

Table 7 BERT Trimodal Analysis

Emotion	Top 20 Features	Audio	Video	Text	Dominant Modality
Fear	A1, A12, V4, T564, V34, T763, V17, T370, V7, T500, T691, V9, T552, T606, T122, T720, T563, T68, T60, T555	10%	25%	65%	TEXT
Sad	A1, A12, V7, V9, V4, V6, V8, V10, V5, V13, V34, V19, A63, V2, A20, A33, T323, A34, V22, V24	30%	65%	5%	VIDEO
Angry	A1, V4, V1, A12, V10, V9, V7, V19, V24, T529, T79, V8, V21, T142, T11, T370, V17, T713, T524, V15	10%	55%	35%	VIDEO
Disgust	A1, V3, V17, V34, T125, T161, T291, V5, V18, T43, T402, T471, T552, T275, T547, V7, T121, T340, T220, T440	5%	30%	65%	TEXT
Happy	A1, V4, V9, A12, V7, V19, V1, V6, V10, V15, V29, V20, V16, T206, T162, V23, T215, T566, T44, V25	10%	65%	25%	VIDEO
Surprise	A1, V8, T12, V14, T308, T340, T694, T402, T165, T685, T337, A12, T496, T754, T627, T46, T257, T587, T711, T498	10%	10%	80%	TEXT

On comparing the weighted accuracies obtained by the Top 20 Features Models with the baseline GraphMFN Model which uses all 409 features[3], we observed that three of our reduced feature models BERT Bimodal(T,A), BERT Bimodal(T,V) as well as the BERT Trimodal models were able to outperform the baseline for the Disgust emotion class. Four of our Top Twenty Feature models - BERT Bimodal(T,A), Bimodal(T,V), BERT Trimodal, Bimodal(A,V) models were able to outperform the baseline for the Happy emotion class. Overall, our reduced feature models were able to obtain better weighted accuracies than the baseline for 2 out of 6 emotions.

Table 8 All Features - F1 Scores

Model	Fear	Sad	Angry	Disgust	Happy	Surprise	Average
Baseline[3]	89.90	66.90	72.80	76.60	66.30	85.50	76.33
GloVe (T,A)	87.77	66.14	67.02	74.88	60.63	85.99	73.74
BERT (T,A)	87.90	72.40	73.80	84.10	69.00	86.70	78.98
GloVe (T,V)	87.77	66.21	67.82	75.29	65.49	85.98	74.76
BERT (T,V)	87.80	72.40	73.60	84.10	71.60	86.60	79.35
GloVe (A,V,T)	87.77	67.55	67.71	74.88	66.57	86.05	75.09
BERT (A,V,T)	87.80	72.30	73.90	84.20	71.80	86.60	79.43
Bimodal (A,V)	87.80	67.80	70.00	76.00	67.60	86.00	75.87

Table 9 Top Twenty Features - F1 Scores

Model	Fear	Sad	Angry	Disgust	Happy	Surprise	Average
GloVe (T,A)	87.77	65.34	67.02	74.88	53.44	85.99	72.41
BERT (T,A)	87.74	71.18	75.89	84.32	69.09	86.62	79.14
GloVe (T,V)	87.77	65.30	67.15	75.24	65.74	85.99	74.53
BERT (T,V)	87.79	66.35	74.91	84.27	71.28	86.38	78.49
GloVe (A,V,T)	87.77	65.56	67.02	74.88	66.25	85.99	74.58
BERT (A,V,T)	87.78	67.11	74.27	83.99	70.36	86.58	78.35
Bimodal (A,V)	87.79	67.58	68.81	76.25	67.75	86.05	75.70

Table 10 All Features - Weighted Accuracy (%)

Model	Fear	Sad	Angry	Disgust	Happy	Surprise	Average
Baseline[3]	62.00	60.40	62.0	69.10	66.30	53.70	62.35
GloVe (T,A)	50.00	50.69	50.00	50.00	60.29	50.00	51.83
BERT (T,A)	50.20	58.50	57.60	72.30	68.70	52.10	59.90
GloVe (T,V)	50.00	50.77	50.64	50.52	65.14	50.04	52.85
BERT (T,V)	50.20	58.70	57.30	72.20	71.30	51.90	60.27
GloVe (A,V,T)	50.00	51.94	50.62	50.00	66.61	50.18	53.22
BERT (A,V,T)	50.20	58.50	57.70	72.30	71.40	51.90	60.33
Bimodal (A,V)	50.00	52.10	52.80	51.40	67.30	50.00	53.93

Table 11 Top Twenty Features - Weighted Accuracy (%)

Model	Fear	Sad	Angry	Disgust	Happy	Surprise	Average
GloVe (T,A)	50.00	50.03	50.00	50.00	56.09	50.00	51.02
BERT (T,A)	49.94	56.47	61.75	72.76	68.76	51.71	60.23
GloVe (T,V)	50.00	50.00	50.11	50.44	65.39	50.00	52.66
BERT (T,V)	50.00	50.83	59.12	71.85	71.36	51.08	59.04
GloVe (A,V,T)	50.00	50.20	50.00	50.00	66.12	50.00	52.72
BERT (A,V,T)	50.06	51.48	59.34	73.08	69.97	51.54	59.24
Bimodal (A,V)	50.00	51.88	51.59	51.85	67.65	50.00	53.83

Overall, the reduced feature models curated by using the top twenty features obtained by interpreting the all feature list using the GradientSHAP algorithm were able to achieve weighted accuracies and F1 scores which were comparable to the ones achieved by their corresponding all feature

models. In some scenarios, especially for the angry, happy and disgust classes, the top twenty feature model even achieved better performance than our all-feature models. They also outperformed the baseline GraphMFN model [3] in 4 out of 6 emotion classes in terms of F1 scores and 2 out of 6 emotion classes in terms of weighted accuracies.

5 | CONCLUSION

This study throws light on the hazards posed by using black-box AI/deep-learning models for critical tasks in Trustworthy systems like Emotion Recognition and explains the importance of making these models explainable/interpretable to humans. It focuses on interpreting the importance of individual features from various modalities (audio, video, text) in Context-Aware Multimodal Emotion Recognition. In the process, we highlight the problems of using pre-extracted anonymous features and employ a relevant XAI method called GradientSHAP for evaluating these features. The XAI method implementation leads to finding a subset consisting of Top Twenty Features for various models described in section 4.1. We then compare the performance results of these lighter, reduced feature models in terms of F1 scores and weighted accuracies with their corresponding all feature models as well as the baseline model GraphMFN [3]. The results show that these smaller models with the advantage of being lighter to train and test. achieve comparable results to their all-feature counterparts and even outperform some of them. They also outperform the baseline model in 4 out of 6 emotion classes in terms of F1 scores and 2 out of 6 emotion classes in terms of weighted accuracies.

One of the limitations of this study is the use of a post-hoc XAI method for interpretability. Even though Model-agnostic (or post-hoc) methods allow interpreting complex machine learning models without understanding their underlying mechanism, they do face the challenge of striking a balance between flexibility and interpretability [65].

The success of these reduced feature models suggests that employing XAI methods to interpret black-box deep-learning models can help us to carefully select high-quality, highly contributing features that can help curate trustworthy AI systems. It is hoped that this research will contribute to a deeper understanding of evaluating significant features using explainable methods (rather than blindly using all pre-extracted features) and unlock their potential to improve the performance and robustness of the system.

ACKNOWLEDGMENTS

Financial disclosure

- The authors did not receive support from any organization for the submitted work.
- No funding was received to assist with the preparation of this manuscript.
- No funding was received for conducting this study.
- No funds, grants, or other support was received.

Conflict of interest

- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no competing interests to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.

APPENDIX

Feature Contribution(s) based on Gradient SHAP values

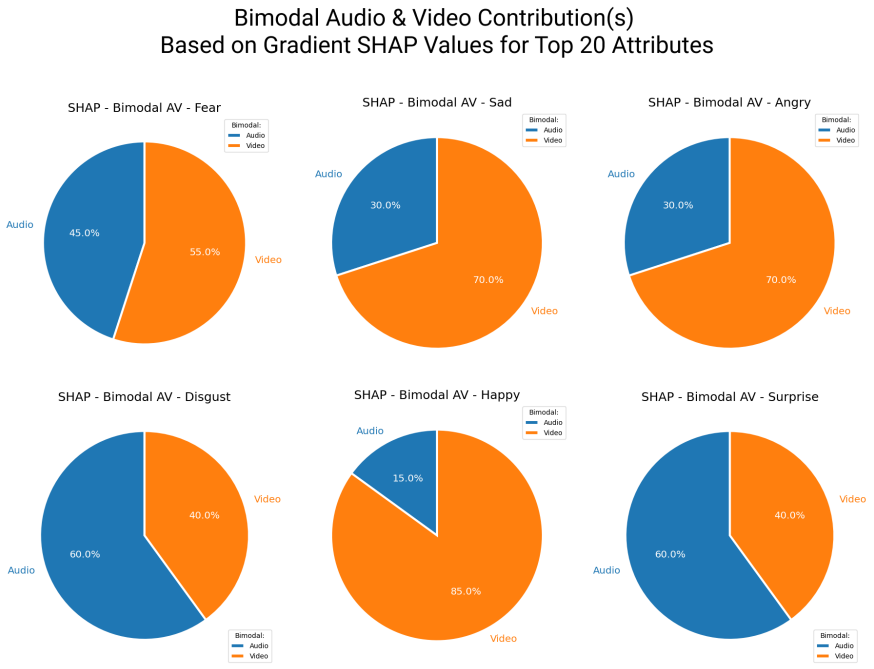

Figure 9 Bimodal Audio and Video Contribution(s) Based on Gradient SHAP Values for Top 20 Attributes

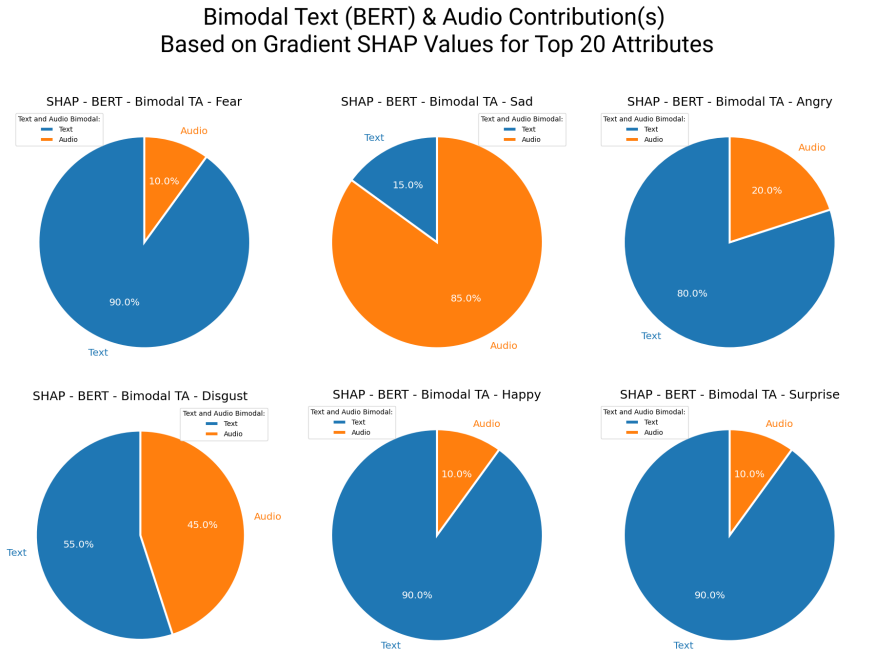

Figure 10 Bimodal Text (BERT) and Audio Contribution(s) Based on Gradient SHAP Values for Top 20 Attributes

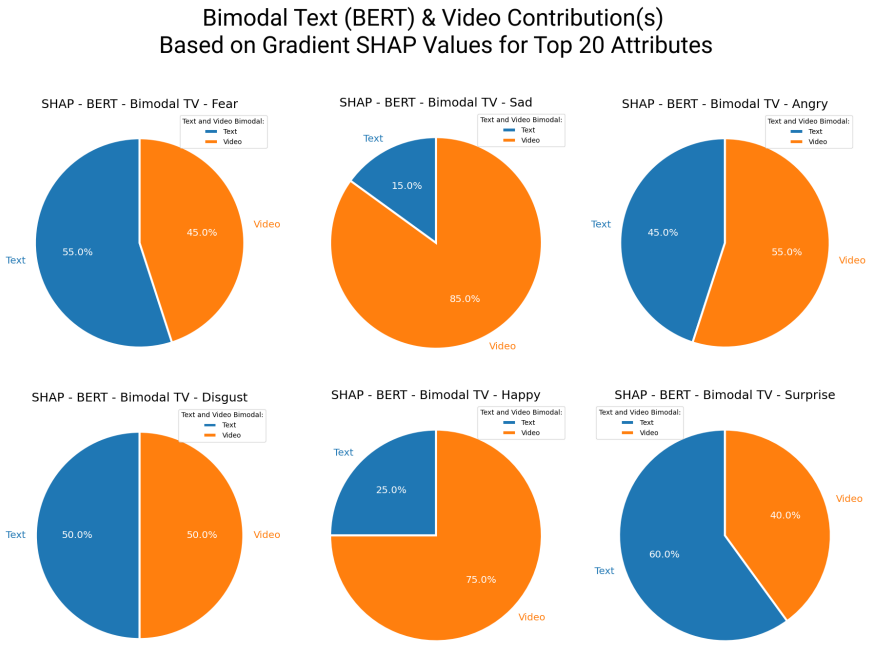

Figure 11 Bimodal Text (BERT) and Video Contribution(s) Based on Gradient SHAP Values for Top 20 Attributes

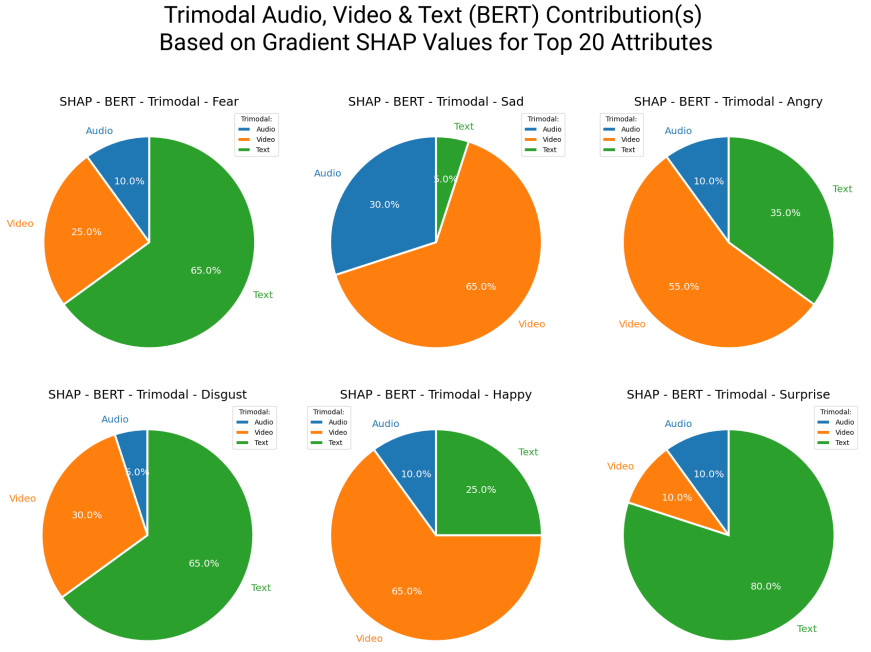

Figure 12 Trimodal Audio, Video and Text (BERT) Contribution(s) Based on Gradient SHAP Values for Top 20 Attributes

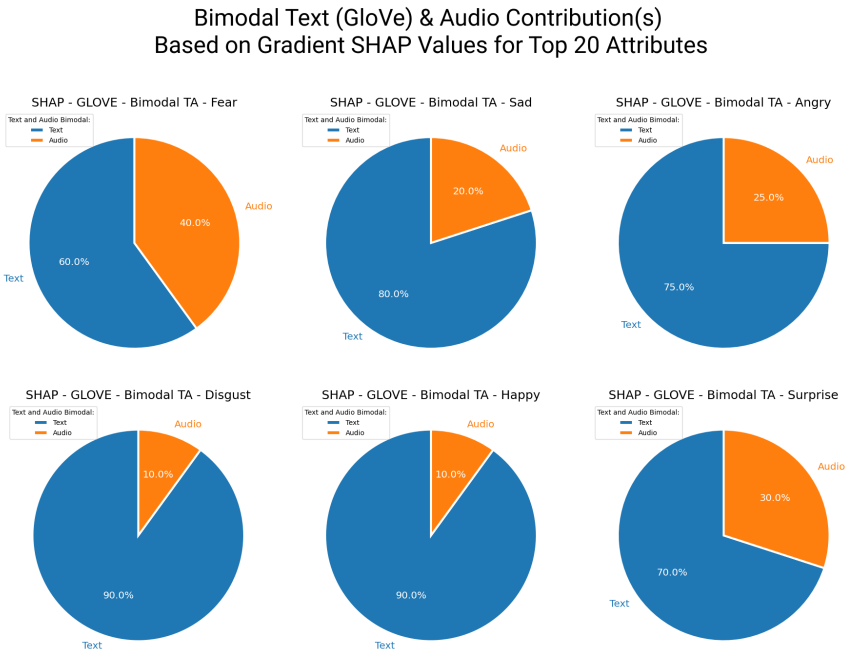

Figure 13 Bimodal Text (GloVe) and Audio Contribution(s) Based on Gradient SHAP Values for Top 20 Attributes

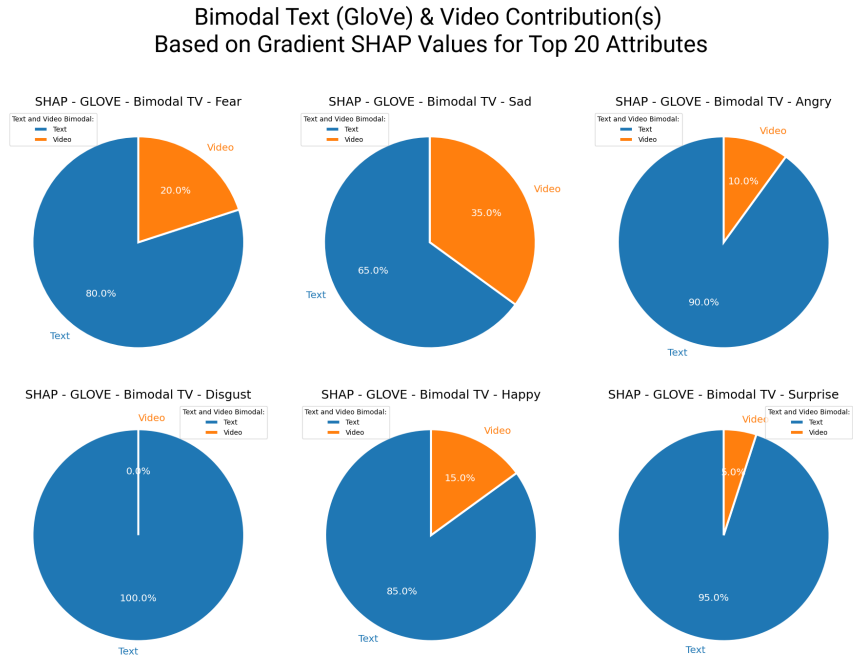

Figure 14 Bimodal Text (GloVe) and Video Contribution(s) Based on Gradient SHAP Values for Top 20 Attributes

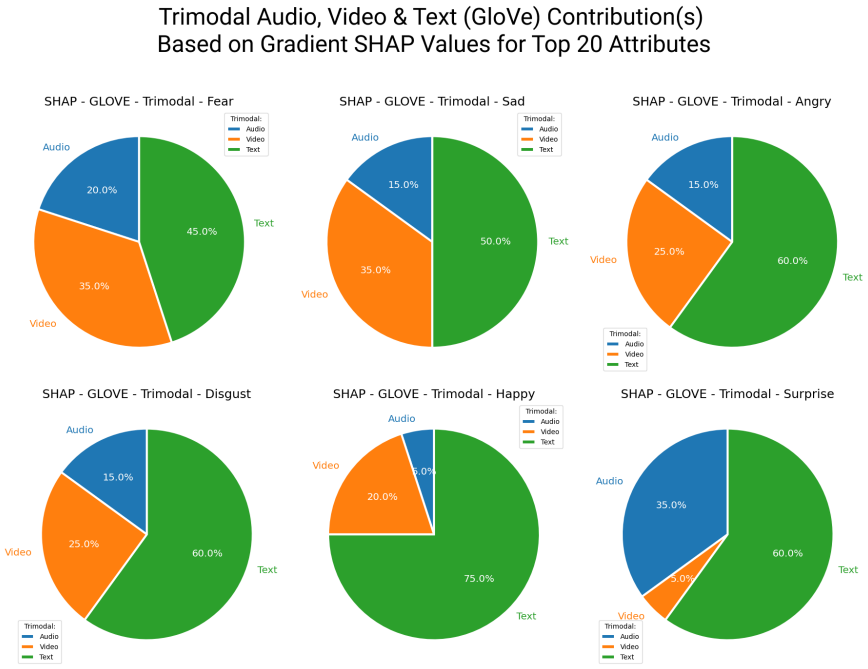

Figure 15 Trimodal Audio, Video and Text (GloVe) Contribution(s) Based on Gradient SHAP Values for Top 20 Attributes

References

1. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 2017; 37: 98–125. doi: 10.1016/j.inffus.2017.02.003

2. Jiang Y, Li W, Hossain MS, Chen M, Alelaiwi A, Al-Hammadi M. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion* 2020; 53: 209–221. doi: 10.1016/j.inffus.2019.06.019

3. Zadeh A, Liang P, Vanbriesen J, et al. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. 2018: 2236–2246.

4. Kumar A, Vepa J. Gated Mechanism for Attention Based Multi Modal Sentiment Analysis. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2020. doi: 10.1109/icassp40776.2020.9053012

5. Verma S, Wang J, Ge Z, et al. Deep-HOSeq: Deep Higher Order Sequence Fusion for Multimodal Sentiment Analysis. *2020 IEEE International Conference on Data Mining (ICDM)* 2020. doi: 10.1109/icdm50108.2020.00065

6. Zhang D, Li S, Zhu Q, Zhou G. Multi-Modal Sentiment Classification With Independent and Interactive Knowledge via Semi-Supervised Learning. *IEEE Access* 2020; 8: 22945–22954. doi: 10.1109/access.2020.2969205

7. Tsai YHH, Ma M, Yang M, Salakhutdinov R, Morency LP. Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2020. doi: 10.18653/v1/2020.emnlp-main.143

8. Liang T, Lin G, Feng L, Zhang Y, Lv F. *Attention is not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion*.

9. Shenoy A, Sardana A. Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)* 2020. doi: 10.18653/v1/2020.challengehml-1.3

10. Hagras H. Toward Human-Understandable, Explainable AI. *Computer* 2018; 51(9): 28–36. doi: 10.1109/mc.2018.3620965

11. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 2018; 51(5): 1–42. doi: 10.1145/3236009

12. Turek M. DARPA - Explainable Artificial Intelligence (XAI) Program. 2017.

13. Rai A. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science* 2019; 48(1): 137–141. doi: 10.1007/s11747-019-00710-5

14. Molnar C. *Interpretable Machine Learning*. 2 ed. 2022.

15. Du M, Liu N, Hu X. Techniques for Interpretable Machine Learning. 2019.

16. Das A, Rad P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. 2020.

17. Bennetot A, Donadello I, Qadi AE, et al. A Practical Tutorial on Explainable AI Techniques. 2021.

18. Qi H, Ece X, Hall F, Iyengar S, Chakrabarty K, Hall H. Multisensor Data Fusion in Distributed Sensor Networks Using Mobile Agents. *Proceedings of International Conference of Information Fusion* 2000.

19. A2Zadeh . A2Zadeh/CMU-MultimodalSDK: CMU MultimodalSDK is a machine learning platform for development of advanced multimodal models as well as easily accessing and processing multimodal datasets.. 2018.

20. Khalane A, Shaikh T. Context-Aware Multimodal Emotion Recognition. In: Springer Singapore; 2021. In press

21. Yang K, Xu H, Gao K. *CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis*: 521–528; New York, NY, USA: Association for Computing Machinery . 2020.

22. Ho NH, Yang HJ, Kim SH, Lee G. Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network. *IEEE Access* 2020; 8: 61672–61686. doi: 10.1109/access.2020.2984368

23. Han W, Chen H, Gelbukh A, Zadeh A, Morency Lp, Poria S. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. *Proceedings of the 2021 International Conference on Multimodal Interaction* 2021. doi: 10.1145/3462244.3479919

24. Lee S, Han DK, Ko H. Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification. *IEEE Access* 2021; 9: 94557–94572. doi: 10.1109/access.2021.3092735

25. Yang B, Shao B, Wu L, Lin X. Multimodal sentiment analysis with unidirectional modality translation. *Neurocomputing* 2022; 467: 130-137. doi: 10.1016/j.neucom.2021.09.041

26. Li Z, Guo Q, Feng C, et al. Multimodal Sentiment Analysis Based on Interactive Transformer and Soft Mapping. *Wireless Communications and Mobile Computing* 2022; 2022. doi: 10.1155/2022/6243347

27. Lian Z, Liu B, Tao J. SMIN: Semi-supervised Multi-modal Interaction Network for Conversational Emotion Recognition. *IEEE Transactions on Affective Computing* 2022. doi: 10.1109/TAFFC.2022.3141237

28. Abdu S, Yousef A, Salem A. Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey. *Information Fusion* 2021; 76: 204-226. doi: 10.1016/j.inffus.2021.06.003

29. Han W, Chen H, Gelbukh A, Zadeh A, Morency LP, Poria S. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In: ; 2021: 6-15

30. Shen J, Zheng J, Wang X. MMTrans-MT: A Framework for Multimodal Emotion Recognition Using Multitask Learning. In: ; 2021: 52-59

31. Mittal T, Bera A, Manocha D. Multimodal and Context-Aware Emotion Perception Model with Multiplicative Fusion. *IEEE Multimedia* 2021; 28(2): 67-75. doi: 10.1109/MMUL.2021.3068387

32. Park J, Kim MH, Choi DG. Correspondence learning for deep multi-modal recognition and fraud detection. *Electronics (Switzerland)* 2021; 10(7). doi: 10.3390/electronics10070800

33. Khare A, Parthasarathy S, Sundaram S. Self-Supervised Learning with Cross-Modal Transformers for Emotion Recognition. In: ; 2021: 381-388

34. Yan X, Xue H, Jiang S, Liu Z. Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling. *Applied Artificial Intelligence* 2021. doi: 10.1080/08839514.2021.2000688

35. Ito K, Fujioka T, Sun Q, Nagamatsu K. Audio-Visual Speech Emotion Recognition by Disentangling Emotion and Identity Attributes. In: . 1. ; 2021: 596-600

36. Georgiou E, Paraskevopoulos G, Potamianos A. M3: Multimodal masking applied to sentiment analysis. In: . 1. ; 2021: 511-515

37. Qi Q, Lin L, Zhang R. Feature extraction network with attention mechanism for data enhancement and recombination fusion for multimodal sentiment analysis. *Information (Switzerland)* 2021; 12(9). doi: 10.3390/info12090342

38. Lee S, Han D, Ko H. Multimodal Emotion Recognition Fusion Analysis Adapting BERT with Heterogeneous Feature Unification. *IEEE Access* 2021; 9: 94557-94572. doi: 10.1109/ACCESS.2021.3092735

39. Huddar M, Sannakki S, Rajpurohit V. Attention-based multi-modal sentiment analysis and emotion detection in conversation using rnn. *International Journal of Interactive Multimedia and Artificial Intelligence* 2021; 6(6): 112-121. doi: 10.9781/ijimai.2020.07.004

40. Sun J, Yin H, Tian Y, Wu J, Shen L, Chen L. Two-Level Multimodal Fusion for Sentiment Analysis in Public Security. *Security and Communication Networks* 2021; 2021. doi: 10.1155/2021/6662337

41. Verma S, Wang J, Ge Z, et al. Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis. In: . 2020-November. ; 2020: 561-570

42. Yang K, Xu H, Gao K. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. In: ; 2020: 521-528

43. Qureshi S, Dias G, Hasanuzzaman M, Saha S. Improving Depression Level Estimation by Concurrently Learning Emotion Intensity. *IEEE Computational Intelligence Magazine* 2020; 15(3): 47-59. doi: 10.1109/MCI.2020.2998234

44. Li H, Tu M, Huang J, Narayanan S, Georgiou P. Speaker-invariant affective representation learning via adversarial training. In: . 2020-May. ; 2020: 7144-7148

45. Kumar A, Vepa J. Gated Mechanism for Attention Based Multi Modal Sentiment Analysis. In: . 2020-May. ; 2020: 4477-4481

46. Li X, Chen M. Multimodal Sentiment Analysis with Multi-perspective Fusion Network Focusing on Sense Attentive Language. In: ; 2020: 1089-1100.

47. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D. M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In: ; 2020: 1359-1367.

48. Siriwardhana S, Reis A, Weerasekera R, Nanayakkara S. Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition. In: . 2020-October. ; 2020: 3755-3759

49. Khare A, Parthasarathy S, Sundaram S. Multi-modal embeddings using multi-task learning for emotion recognition. In: . 2020-October. ; 2020: 384-388

50. Li JL, Lee CC. Using speaker-aligned graph memory block in multimodally attentive emotion recognition network. In: . 2020-October. ; 2020: 389-393

51. Li B, Li C, Duan F, Zheng N, Zhao Q. TPFN: Applying Outer Product Along Time to Multimodal Sentiment Analysis Fusion on Incomplete Data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2020; 12369 LNCS: 431-447. doi: 10.1007/978-3-030-58586-0$_26$

52. Li X, Chen M. Multimodal Sentiment Analysis with Multi-perspective Fusion Network Focusing on Sense Attentive Language. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2020; 12522 LNAI: 359-373. doi: 10.1007/978-3-030-63031-7$_26$

53. Chen F, Luo Z, Xu Y, Ke D. Complementary fusion of multi-features and multi-modalities in sentiment analysis. In: . 2614. ; 2020: 82-89.

54. Ho NH, Yang HJ, Kim SH, Lee G. Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network. *IEEE Access* 2020; 8: 61672-61686. doi: 10.1109/ACCESS.2020.2984368

55. Ghosal D, Akhtar M, Chauhan D, Poria S, Ekbal A, Bhattacharyya P. Contextual inter-modal attention for multi-modal sentiment analysis. In: ; 2020: 3454-3466.

56. Chandra E, Hsu JJ. Deep Learning for Multimodal Emotion Recognition-Attentive Residual Disconnected RNN. In: ; 2019

57. Dumpala S, Sheikh I, Chakraborty R, Kopparapu S. Sentiment Classification on Erroneous ASR Transcripts: A Multi View Learning Approach. In: ; 2019: 807-814

58. Akhtar M, Chauhan D, Ghosal D, Poria S, Ekbal A, Bhattacharyya P. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In: . 1. ; 2019: 370-379.

59. Sangwan S, Chauhan D, Akhtar M, Ekbal A, Bhattacharyya P. Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis. *Communications in Computer and Information Science* 2019; 1142 CCIS: 662-669. doi: 10.1007/978-3-030-36808-1$_72$

60. Bagher Zadeh A, Liang PP, Poria S, Cambria E, Morency LP. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In: Association for Computational Linguistics; 2018; Melbourne, Australia: 2236–2246

61. Lundberg S, Allen P, Lee SI. *A Unified Approach to Interpreting Model Predictions*.

62. Knapič S, Malhi A, Saluja R, Främling K. Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain. *Machine Learning and Knowledge Extraction* 2021; 3(3): 740–770. doi: 10.3390/make3030037

63. Kokhlikyan N, Miglani V, Martin M, et al. Captum: A unified and generic model interpretability library for PyTorch. 2020.

64. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. 2018.

65. Ribeiro MT, Singh S, Guestrin C. Model-Agnostic Interpretability of Machine Learning. 2016.