Multimedia Tools and Applications (2025) 84:14167–14185
https://doi.org/10.1007/s11042-024-20184-0

Multimedia video analytics using deep hybrid fusion algorithm

Jamuna S. Murthy[1] · **G. M. Siddesh**[2]

Received: 19 October 2022 / Revised: 7 September 2023 / Accepted: 27 August 2024 /
Published online: 23 September 2024

Abstract
In recent times, Multimedia Video Analytics is considered as one of the trending areas of research that focus on understanding different modalities of user-generated data. One of the sophisticated approaches to address this area of research has been to develop an efficient fusion technique that can handle multimedia data such as video which is a combination of text, image and speech data. But most of the researchers failed to fill the gap between different modalities as the video signals are heterogeneous, invariant and this poses a major challenge even today. In contrast to previous research, proposed study examines implementation of novel framework for Multimodal Video Analysis called "Multimedia Video Analytics using Deep Hybrid Fusion Algorithm" which is a smart video analytical framework with three major modules such as Video Feature Extraction, Modality Representation and Fusion. Firstly the three modalities of video such as text, image and speech are extracted and represented in subspace as six hidden vectors using deep learning approach called “Modality Unchanged Precise Representation” or MUR Algorithm which uses Encoder Decoder Representation of BiLSTM. Later a novel video fusion technique called Deep Hybrid Fusion algorithm built over Attention based Transformation technique using Softmax suppression is used to fuse the six hidden vectors in subspace for further task prediction. The proposed DHF approach is compared against fusion variants of LSTM such as MFN, TFN, MRN, MRMF, MV-LSTM and applied for humor detection task on classic video datasets such as IEMOCAP, CMU-MOSI, CMU-MOSEI. By using metrics such as Precision, Recall, F-Measure and Accuracy the proposed DHF algorithm outperformed to provide best 7 class accuracy of 95.84%.

Keywords Humour Detection · Multimodal · Unchanged · Precise · Long short term memory (LSTM)

✉ Jamuna S. Murthy
jamunamurthy.s@gmail.com

G. M. Siddesh
siddeshgm14@gmail.com

[1] Department of Computer Science and Engineering, M.S. Ramaiah Institute of Technology (Affiliated to Visvesvaraya Technological University, Belgaum), India

[2] M.S. Ramaiah Institute of Technology (Affiliated to Visvesvaraya Technological University, Belgaum), India

1 Introduction

The extraction of pertinent information from multimedia data, such as video and audio, is the focus of the developing field of Multimedia Video Analytics (MVA). This method is used to analyse the data and derive important insights that have applications in a variety of industries, such as security, surveillance, and entertainment. The most recent studies in the subject of multimodal video analytics from 2022 have suggested a number of novel approaches to video analysis [1–5] and some of them are explained here.

In this study [6] on graph Convolutional networks (GCNs), a new technique for multimodal semantic video analysis, are proposed. The developed algorithm can examine both the textual and visual content of movies, modelling the interactions between various modalities using GCNs. The experiments' findings demonstrate that the proposed method performs better than current methods in terms of precision and effectiveness.

In this research [7] on a multimedia analytics based on deep learning analysis of video footage is presented for the users. The suggested technique extracts features from films and analyses their content using Convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The results of the studies demonstrate that the suggested method performs video content analysis tasks with high accuracy and efficiency.

A hybrid method for object recognition and semantic segmentation in multimedia video analytics was presented in 2022 by these researchers [8]. The proposed technique combines the advantages of both methodologies to produce precise and effective video analysis. The experiments' findings demonstrate that the suggested approach exceeds current approaches in terms of precision and effectiveness.

For the purpose of recognizing actions in videos, this researcher [9] put up a spatiotemporal attention network. In order to selectively concentrate on crucial spatio-temporal regions in films for action recognition, the suggested method makes use of attention mechanisms. The experiments' findings demonstrate that the proposed method performs better than current methods in terms of precision and effectiveness.

Here [10], a multi-task learning strategy is used to track and recognize objects in films. Using a single deep neural network, the suggested method jointly learns object identification and tracking tasks. The results of the studies demonstrate that the suggested method performs object detection and tracking tasks with high accuracy and efficiency.

Overall, new methods for analyzing videos have been offered by current research in the subject of multimedia video analytics. Deep learning methods including CNNs, RNNs, GCNs, and attention mechanisms are extensively used in the previous methods. In a variety of video analysis tasks, including object detection, tracking, action recognition, and content analysis, these approaches have demonstrated promising results in terms of accuracy and speed.

Analyzing and comprehending vast amounts of multimedia data is the difficult research topic of multimedia video analytics. Among the MVA research difficulties now being addressed include real-time processing, deep learning for MVA, multimodal analysis, explainable MVA, efficient fusion techniques, privacy and security. Although some approaches have been suggested in recent research publications as solutions to these problems, it is still necessary to develop MVA techniques for large-scale multimedia data that are more efficient and effective.

The first research challenge analyzed in the proposed research work is related to multimodal analysis of video data. Multimodal analysis is a kind of analysis that takes into account a wide variety of sentiments, including audio, video, and text. Similar to the conventional sentimental analysis, multimodal sentimental analysis is based on the same principles. Multimodal sentiment analysis has been used recently in the development of

multimodal recommender systems. Learning about the user's experience, which is communicated through multimodal analysis, is the key or the lead to recommender systems. It is possible to produce more effective insights with multimodal sentiment analysis.

Due to the expanding availability of video data and the possibility of deep learning algorithms for analysing and comprehending video content, the field of research into video modality representation is expanding quickly. Although there have been numerous important developments in video modality representation over the past few years, there are still a number of research issues that need to be resolved and this is ext research challenge addressed in the proposed work.

Although the study of video modal representation is advancing quickly, there are still a number of problems that need to be solved. Creating efficient representation learning methods for video data, domain adaptation, multi-modal fusion, and explainability and interpretability are a few of the major obstacles. In order to advance the state-of-the-art in video modality representation and realise the full potential of video data for various applications, it will be essential to address these problems.

Due to their capacity to combine data from many sources, video fusion techniques have drawn a lot of attention in recent years. This has improved the quality of the visual content. To overcome the shortcomings of current approaches, brand-new video fusion techniques have been presented. In recent years, numerous inventive video fusion approaches have been put forth. A deep learning-based video fusion technique was put out by Zhang et al. (2022) in their work to enhance the detection of objects in low-light conditions by combining RGB, depth, and thermal pictures in real-time [10].

A multi-scale Convolutional neural network was suggested in a different study by Liu et al. (2022) [7] for combining dynamic visible and infrared films, which enhances the detection of moving objects in cluttered backgrounds. A hybrid video fusion method was put forth in a recent work by Chen et al. (2023) [9], which merged methods based on deep learning and sparse representation to produce better results in terms of image quality and fusion accuracy.

Although video fusion techniques have advanced, there are still a number of problems that need to be solved. Below are some of these challenges discussed:

1. **Video fusion that is dynamic:** Video fusion that is dynamic is a difficult problem, especially when the videos have various frame rates and resolutions. To overcome this obstacle and enhance the detection of moving objects in cluttered backgrounds, novel fusion approaches are needed.
2. **Computational complexity:** Deep learning-based video fusion techniques are computationally expensive, which limits their practical applicability. Developing novel fusion techniques that are computationally efficient, without sacrificing fusion accuracy, is an ongoing research challenge.

In order to develop more sophisticated video fusion techniques with higher performance and practical applicability, these issues must be resolved and hence the current research work focus on developing a novel video fusion strategy for videos as final research challenge. Thus the novelty of the proposed research work involves:

i. Creating a *smart multimedia video analytical framework* to handle the research problem of multimedia video feature extraction of modalities such as text, image and audio using deep learning techniques.
ii. To put into practise a *novel video modality representation algorithm* to address issues with domain adaptation, explainability and interpretability and efficient

representation learning using advanced neural networks and BiLSTM Encoder-Decoder techniques.

iii. To propose a *novel multimodal fusion algorithm* that addresses current difficulties with fusing the image, text and audio data over real-time that reduces computational complexity and allows dynamic fusion.

In contrast to previous research, this study examines implementation of novel framework for Multimodal Video Analysis called "Multimedia Video Analytics using Deep Hybrid Fusion Algorithm" which is a smart video analytical framework with three major modules such as Video Feature Extraction, Modality Representation and Fusion.

The technique of locating and extracting significant patterns and characteristics from video data is known as "video feature extraction" i.e. done by Video Feature Extraction module of the framework. To glean insights from the video information, this module entails analyzing a variety of modalities, including text, image, and speech. Proposed Enhanced-BERT model is to analyse the textual information in the video in order to find and extract significant keywords and phrases. During image feature extraction, crucial aspects like color, texture, and shape are recognized and extracted from the visual content of the video using proposed neural network algorithm called MultiNet-101. In speech feature extraction, critical information like pitch, loudness, and tone are extracted from the audio material of the video using signal processing techniques and Librosa Library.

Next Modality Representation module introduces a novel Video Modality Representation strategy called "Modality Unchanged-Precise Representation" which represents the extracted text, image and speech modalities in subspace using six hidden vectors.

Influenced by the latest advancements in algorithms that use more than one source of the domain, proposed Modality Unchanged-Precise Representation technique learns to represent each modality clearly. The first representation is focused on lessening the modality gap. Here, all the modalities for input are delegated to a commonplace. Multimodal inputs are taken from various sources even though they serve a common goal of the orator. The unvaried input helps to get these fundamental resemblances and their corresponding features. Many of the older sentiment analysis models do not use arrangements like this before integrating various sources of data, which adds additional work on them later, while they want to integrate different modes of data.

Affixing to the unvaried subcategory, proposed technique additionally gains knowledge about Modality Unchanged characteristics that are specific to all the modalities. For the input, each modality clasps the different characters that are well defined, which include speaker-centric data. Such individual data are often not connected to further modalities and are classified as disturbances. These can be of the form where the speaker's propensity is sarcastic, ironic or sometimes doubtful. Gaining knowledge about such modality-centric characteristics accolades the inherent features apprehended in the Precise or Unvarying space and will provide a complete multimodal portrayal of the inputs.

After representing the modalities in the subspace a novel fusion algorithm called "Deep Hybrid Fusion" Algorithm is introduced to fuse the hidden vectors represented in the subspace using Modality Unchanged-Precise Representation using attention-based transformation method using Softmax suppression.

Later to gain knowledge about these subspaces, a mixture of various types of droppings or loss functions are introduced which includes orthogonal loss, reconstruction loss, and distributional similarity loss. The proposed framework is applied for humour detection and tested against classic datasets such as IEMOCAP, CMU-MOSI, CMU-MOSEI using metrics precision, recall, F-Measure, Accuracy which outperformed by providing best accuracy of 95.84%.

2 Related works

Multimedia Video Analytics (MVA) is a field of study that involves the processing and analysis of multimedia data, particularly video data. MVA is used in various applications, such as surveillance, security, and entertainment. In recent years, there has been a significant increase in the research and development of MVA techniques, particularly with the advancement of deep learning and computer vision technologies. This literature review aims to provide a comprehensive overview of the latest research in MVA using research papers from the recent years.

2.1 Multimedia data analytics

The composition of proposed framework is mainly classified into (i) input level (ii) context level. While input-level algorithms deal with an ideal input in separation, contextual algorithms use bordering inputs from the all-inclusive video.

- **Input level:** The work in this division is mainly concentrated on cross-modal learning that uses sophisticated fusion methods. These works incorporate a collection of methods, such as tensor-based learning [11] and multi-kernel learning [12].When these works achieve fusion on the representation of the input, another division of work performs fusion on the world level. The Approaches that are used are multimodal-aware word implants [13], reiterative multi-stage fusion [14], graph-based fusion [15, 16] and reiterative neural networks [17–19].
- **Context level:** The context from neighbouring input values of the target input is used in these models. Constructed as a hierarchical lattice, they model discrete inputs at the subordinate level and context inputs at the next plane. Sherstinsky, A et al. put forward distinct initial models, bi-directional LSTMs [20], which used this blueprint towards bc-LSTM for the context depiction learning, mounting the comprehensive complication as an organized prognosis work [21]. Later tasks involve creating improved contextual models. The work we do is very unalike from these obtainable works. We neither abstain from using contextual data nor concentrate on complex fusion methods. Rather, we work on the consequence of representation learning before fusion. However, proposed model can include these above specified factors, if needed [22–24].

2.2 Subspace representations in multimedia data analytics

Generally works based on cross-modal common subspace are classified into 3 types. The first type of model is Correlation based Algorithms [25] which learn cross-modal correlations using different correlation techniques [26]. The second type of model is the one that learns from a modality mapped subspace representations which are shared at the same time using methods such as adversarial learning [27]. The third type of model is the one that translates from one modality to another by using different sequence methods [28] and cyclic translation methods [29].

2.3 Factorized representation

Inside the subspace learning system, we focus on factorized depictions. The main motive is to learn modality-invariant and explicit portrayals. We have taken encouragement from various

articles on shared-private representation to accomplish this goal. The idea of shared private representation is mainly taken from the multiview component analysis. The literature works mentioned above have helped us to design Latent Based Models which include private, separate, and shared latent variables. Our proposal, unlike these models, demands a differentiable Deep Neural Network workflow that avoids the need for approximate inference. The proposed framework is mainly related to the Domain Separation Network which focuses on separating two domains.

3 Multimedia video analytical framework

Proposed research investigates the application of an innovative framework for Multimodal Video Analysis for humor detection is videos as shown in Fig. 1. The framework consists of three modules: Video Feature Extraction, Modality Representation and Deep Hybrid Fusion. The Video Feature Extraction module is responsible for identifying and extracting important patterns and attributes from video data. This process involves analyzing various modalities, such as text, images, and audio. The Modality Representation module employs a novel method known as "Modality Unchanged-Precise Representation" to accurately depict domain representation in higher dimesion for each modality. The Deep Hybrid Fusion module presents a novel approach to fuse the modalities represented in subspace using Softmax suppression.

3.1 Video feature extraction

This module extracts three major features such as text, image and speech using various deep learning techniques.

3.1.1 Text feature extraction

A pre-training language model called BERT (Bidirectional Encoder Representations from Transformers) has considerably enhanced the study of natural language processing (NLP). However, the BERT algorithm suffers a number of difficulties, including:

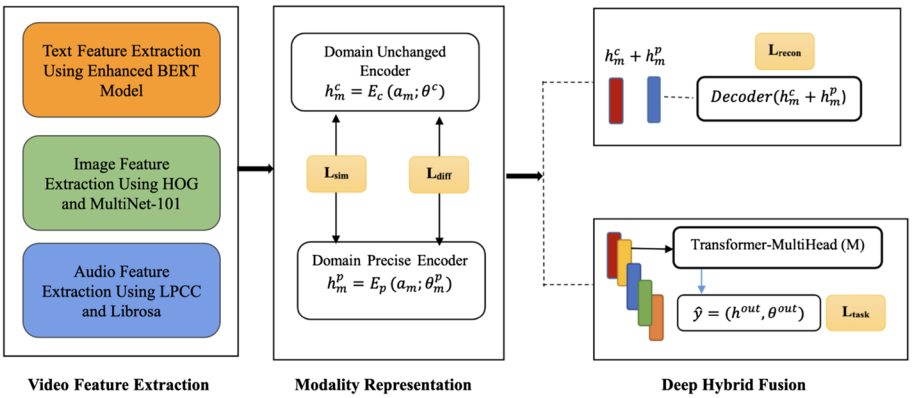

Fig. 1 Proposed framework for Multimedia Video Analytics using Deep Hybrid Fusion Algorithm

1. **Limited comprehension of context:** Despite being a strong NLP tool, BERT still has issues comprehending the context of language. For instance, BERT might not be able to comprehend irony or sarcasm, which could result in inaccurate forecasts.
2. **Training data bias:** BERT, like any machine learning model, is susceptible to bias from the training data that was used to develop it. This can result in incorrect predictions or confirm preexisting prejudices.

Hence in the proposed work an "*Enhanced-BERT*" model is built with more precise training data which can overcome contextual difficulties such as understanding sarcasm, irony that can in turn improvise the accuracy. The proposed methodology for text feature extraction is given by: Firstly input the words into $\mathrm{BERT}_{\mathrm{base}}$ model with Transformer layers, $\mathrm{L}=4$ to average the output. Finally each piece of word is represented as $\mathrm{W_t}=768$ dimensional feature vector $\mathrm{X_t}$.

$$\left\{ \mathrm{X_j^t} \right\}_{j=1}^{M} = BERT_{base}(T) \tag{1}$$

$$\mathrm{X_t} = \frac{1}{L}\left(\sum_{j=1}^{M} \mathrm{X_j^t} \right) \in \mathrm{R^{W_t}} \tag{2}$$

here , $\mathrm{X_j^t}$represents the out of the last jth transformation layer in $\mathrm{BERT}_{\mathrm{base}}$ model for each word T.

3.1.2 Image feature extraction

Out of all the modalities facial features clearly explain the emotion of the person for multimodal analysis. For each video input the emotional change in the facial expression of the person with time series is determined frame by frame and is processed without any background information. Popular library called OpenCV is used for video to frame extraction. Here usually background information of the image doesn't play an important role as the more focus in given on the person's topic of interest or news he is talking about. Next face detection is done by one of the popular libraries Histogram of Oriented Gradients or HOG.

For each video frame $\boldsymbol{V_i}$, the number of faces detected is given by $\boldsymbol{Face_i}$. Let $\boldsymbol{H_i}$ be the height of each face detected from $\boldsymbol{V_i}$. Each face in the frame is of different heights and hence to fill the gap, black block $\boldsymbol{Block_i}$ is used as splicing. This confirms the uniformity of heights of all the faces and also horizontal stitching is applied as final version to the faces in image to input to the neural network. The formulization of the above extraction technique is given in Eq. (3) and Eq. (4).

$$A_{Block_{(Face_i)}} = (3, Length(Face_i), H_i - Height(Face_i)) \tag{3}$$

$$padding(Face_i) = \begin{cases} Face_i, Height(Face_i) = H_i \\ Face_i \oplus Block_i \in R^{X_{Block_{(Face_i)}}}, Height(Face_i) < H_i \end{cases} \tag{4}$$

$$Face_i = stithing(padding(Face_i)) \tag{5}$$

Here,$Length(Face_i)$ is Length of the image, $Height(Face_i)$ I height of the image. Also, $A_{Block_{(Face_i)}}$ is the dimension of the block box and $\oplus$ is the vertical join operator to fuse the different features. Finally $stithing(padding(Face_i))$ represents the padding and horizontal stitching for the incorrect image as shown in Eq. (5).

$Face_i$, is the final stitched image after correction. Next the each image frame is pre-processed to normalize and then fed into MultiNet-101 a novel neural network algorithm with 101 layers. This neural network is pre-trained with 2048 dimensions for feature extraction from $Face_i$. An average value of feature vector $X_I^{Face_i}$ for visual feature extraction is calculated with $W_v = 2048$ dimensions for each "*frame*" of video. The formularization of the feature extraction using MultiNet-101 is given by Eq. (6) & (7) respectively.

$$X_I^{Face_i} = MultiNet - 101(Face_i) \tag{6}$$

$$X_v = \frac{1}{frame}\left(\sum_i X_I^{Face_i}\right) \in R^{W_v} \tag{7}$$

3.1.3 Speech feature extraction

Librosa is a library used for speech extraction in the proposed model. The speech data with time series is inputted to Librosa library with sampling rate of 22,000 Hz. Heuristic based audio extraction technique is used for noise reduction from the sample audios. Next the local features such as MFCC, Spectral Centroid, Mel-spectrogram are extracted from the audios as non over palling windows i.e. $W_{a..}$ A joint vector i.e. $\left\{X_j^a\right\}_{j=1}^{W_a}$ is created by combining all the local features with $W_a = 285$ dimensions. The average value of the join vector is given by:

$$X_j^a = X_i^{MFCC} \oplus X_i^{MFCCdelta} \oplus X_i^{Mel} \oplus X_i^{Meldelta} \oplus X_i^{Spec} \tag{8}$$

$$X_a = \frac{1}{W_t}\left(\sum_{j=1}^{M} X_j^a\right) \in R^{W_a} \tag{9}$$

Here $\oplus$ concatenates each of the features i.e. $X_i^{MFCC}, X_i^{MFCCdelta}, X_i^{Mel}, X_i^{Meldelta}, X_i^{Spec}$ in the Eq. (8) and Eq. (9).

3.2 Modality representation

The Modality Representation module introduces a new method called "Modality Unchanged Precise Representation" Learning. This technique involves mapping different types of data—such as text, images, and speech—into a common space using six hidden vectors. By leveraging recent advancements in algorithms that handle multiple data sources, this approach aims to clearly represent each type of data. The primary goal is to minimize the differences between modalities by aligning them into a shared framework. This alignment helps in identifying common features and similarities among the different data types, making it easier to analyze them together. Traditional models often face challenges when integrating various data sources, leading to additional complexity and effort. The new approach simplifies this process by ensuring that all types of data are effectively combined from the beginning. Additionally, the technique also focuses on identifying characteristics that are consistent across all types of data. It recognizes that each type of data may have unique features or nuances, such as specific traits related to the speaker's tone or style, which might not directly connect with other types of data. Understanding these consistent features enhances the overall analysis by providing a more complete and accurate representation of the combined data.

The MultiModality-Unchanged-Precise Representation is done at two levels firstly the Assertion Level Representation (ALR) and secondly the Multimodality Unchanged-Precise Representations as shown in Algorithm 1:

Algorithm 1 Proposed Multimodality Unchanged-Precise Representation (MUR)

Input: *Set of modalities* m ∈{i, t, s}

Output: *Six modality vectors in common subspace* $\{h_i^p, h_t^p, h_s^p, h_i^c, h_t^c, h_s^c\}$

Begin

for each m ∈{i, t, s}

  map assertion sequence $A_m \in \{i,t,s\}$

    to vector $a_m \in R^{d_h}$

The final BiLSTM (Long Shirt Term Memory) representation using hidden dense layer for Assertion level is given by:

$$a_m = sLSTM(A_m;\ \theta_m^{lstm}\ )$$

for each m ∈{i, t, s}

  map assertion sequence $A_m \in \{i,t,s\}$

    to hidden modality unchanged vector

$$h_m^c = R^{d_h}$$

for each m ∈{i, t, s}

  map assertion sequence $A_m \in \{i,t,s\}$

    to hidden modality precise vector

$$h_m^p = R^{d_h}$$

The final multimodality unchanged-precise representations using Encoding Decoding functions are given by

$$h_m^c = E_c\,(a_m;\theta^c)\ ,\ \ h_m^p = E_p\,(a_m;\theta_m^p)$$

for each modality m ∈{i, t, s}

  generate six hidden vectors $\{h_i^p, h_t^p, h_s^p, h_i^c, h_t^c, h_s^c\}$

    using neural network

here E_c shares θ^c among all three modalities

also E_p assigns θ_m^p parameter to all three modalities

End

Explanation of the algorithm: In step 1 Consider each video data which can be divided into constituent Assertions (i.e. unit of speech bound with breaths and pauses) which in turn are small videos itself.

In step 2 Let *"A"* be represented as the Assertion which constitute three low level features such as *Image (i), Text (t) and Speech (s)* modalities. These low level features are represented as $A_i \in R^{L_i \times D_i}, A_t \in R^{L_t \times D_t}, A_s \in R^{L_s \times D_s}$. Here L_m denotes the length of assertion for the tokens in Feature Dimension D_m for modality *"m"*.

The primary task involved here is to predict the number of Assertions from the sequence $A_m \in \{i,t,s\}$ taken from either a predefined category C with x ∈ R^C or from the Continuous Intensity Variable x∈R.

Firstly Assertion level representation is done by BiLSTM (Long Shirt Term Memory) hidden dense layer using $a_m = sLSTM(A_m;\theta_m^{lstm}$. And secondly Modality Unchanged-Precise Representations are done using Factorized Learning which is a key feature of Domain Separation Network (i.e. a variant of differential Deep Neural Network). Algorithm 1 represents modalities in two subspaces i.e. Modality-Unchanged given by $h_m^c = R^{d_h}$ and Modality-Precise given by $h_m^p = R^{d_h}$.

3.3 Deep hybrid fusion

In order to fuse the hidden vectors represented in the subspace using Modality Unchanged-Precise Representation, a unique fusion algorithm known as "Deep Hybrid Fusion" Algorithm is introduced as shown in Algorithm 2. The attention-based transformation method using Softmax suppression is a powerful tool in the proposed deep hybrid fusion strategy for combining text, image, and speech modalities represented in subspace.

The Softmax function is used in this technique to filter out extraneous or distracting information while focusing attention on specific traits or components of each modality. The outcome is a more trustworthy and accurate representation of the combined modalities, which may then be used for a number of downstream tasks like object recognition, speech recognition, and text categorization. The Softmax suppression helps to increase the discriminative power of the fusion of features by reducing the influence of irrelevant information. Overall, this approach provides a helpful way to combine many modalities to improve performance in a variety of applications.

Algorithm 2 Proposed Deep Hybrid Fusion

Input: *Six modality vectors in common subspace* $\{h_i^p, h_t^p, h_s^p, h_i^c, h_t^c, h_s^c\} \epsilon R^{6\times d_h}$

Output: *Fusion output vector* $h^{out} = [\bar{\bar{h}}_i^p, \oplus \ldots\ldots \oplus \ \bar{h}_s^c \ ] \epsilon R^{6\times d_h}$

Begin

  Initialize input vector as Matrix M=$[h_i^p, h_t^p, h_s^p, h_i^c, h_t^c, h_s^c]$

    and set Q = K = V = M $\epsilon R^{6\times d_h}$

Apply Attention Based Transformer - Softmax suppression

  to generate output

    $\bar{M} = [\bar{h}_i^p, \bar{h}_t^p, \bar{h}_s^p, \bar{h}_i^c, \bar{h}_t^c, \bar{h}_s^c]$

Attention for ith head is given by:

    $\overline{\bar{M} = MultıHead(M;\theta^{att})} = (head_1 \oplus \ldots\ldots \oplus head_n)W^o$

here,

  $\oplus$ - Concatenation and $\theta^{att} = \{W^q, W^k, W^v, W^o\}$

Using transformer output concatenate to get joint vector

  $h^{out} = [\bar{\bar{h}}_i^p, \oplus \ldots\ldots \oplus \ \bar{h}_s^c \ ] \epsilon R^{6\times d_h}$

Finally Task prediction: $\hat{y} = (h^{out}, \theta^{out})$

End

Explanation of algorithm 2: After representing the modalities in two subspaces they are fused using joint vector for final prediction of modalities using Algorithm 2 i.e. DHF algorithm as shown in step 1. The six hidden vectors are fused using Attention based Transformation technique using Softmax suppression as shown in Eq. (10).

$$Attention(Q, K, V) = softmax((QK^t)/(\sqrt{d_h})) \tag{10}$$

where, Q, K, and V are the query, key, and value matrices. The Attention based Transformer computes these kind of similar parallel attention transformers, where each attention output is called a head. The i^{th} head is computed as shown in Eq. (11):

$$head_i = Attention(QW_i^q, KW_i^K, VW_i^V) \tag{11}$$

Here, $W_i^{q/k/v} \epsilon R^{d_h \times d_h}$ are head-precise parameters that project the matrices linearly in to sub spaces.

3.4 Representation of loss function

The regularization of loss is performed by calculating the loss function of Modality Learning and Representation in subspaces done by MUR Algorithm as shown in Eq. (12):

$$\boldsymbol{L = L_{task} + aL_{sim} + \beta L_{diff} + \gamma L_{recon}} \tag{12}$$

Here $\mathbf{L_{task}}$, $\mathbf{L_{sim}}$, $\mathbf{L_{diff}}$, $\mathbf{L_{recon}}$ denote loss functions.

Regularization of loss function L, is carried is determined by interaction weights α, β,. To achieve the desired subspace representation each of the loss function are responsible. Now let's see the different loss function listed above:

- $\mathbf{L_{task}}$**:** The quality of training is depicted by calculation Task Loss function.
- $\mathbf{L_{sim}}$**:** Cross Modality Discrepancy for Modality Unchanged-Precise Representations are calculated using is calculated Similarity Loss.
- $\mathbf{L_{diff}}$**:** Difference Loss ensures that the Modality Unchanged-Precise Representations capture different aspects of the input.
- $\mathbf{L_{recon}}$**:** Reconstruction Loss ensures the hidden representations capture the details of their respective modality.

4 Evaluation results

Here totally three datasets are considered for result evaluation. The descriptions of datasets are given in Table 1 below:

- *IEMOCAP*

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) is a multimodal and multi speaker database collected at SAIL lab at USC. This dataset consists of 10,000 videos annotated with multiple categorical labels such as anger, happy, frustrated, sad etc.

Table 1 Distribution of Benchmark Datasets

Datasets	Name	Videos	Features
D1	IEMOCAP	10,000	Text, Image, Speech
D2	CMU-MOSI	2199	Text, Image, Speech
D3	CMU-MOESI	23500	Text, Image, Speech

In the proposed work this dataset is considered for feature extraction of videos into Text, Image, and Speech and represent them in sub space for modality learning.

- *CMU-MOSI*

This is one of the top datasets which consist of YouTube monologues with different expression of speakers on variety of topics. It has totally 95 videos with 95 distant speakers. It consists of 2199 Assertion-Videos (i.e. short videos) which is taken as input to the for Multimedia Feature extraction. The sample of the dataset is given in Figs. 2, 3 and 4 below.

- *CMU-MOSEI*

This is again an extension to *CMU-MOSI* dataset with 23500 manually annotated Assertion-Videos from 2338 videos with 1000 distant speakers on more than 260 topics. After

Fig. 2 CMU-MOSI dataset with different human expressions

Fig. 3 Gaze directions of CMU-MOSI dataset

Fig. 4 Camera angles with 0°,45°, 90°,135° and 180º

feature extraction, we train the modal for modality representation using MUR Algorithm and finally fuse them using DHF algorithm for task prediction i.e. Humour detection. Here we calculated seven class sentiments.

4.1 Multimedia video analytics with humour detection

Multimodal Video Analytics with Humour Detection for seven classes (i.e. Anger, Disgust, Frustrated, Joyful, Neutral, Surprise and Contempt) is considered as one of the crucial or application part of the research work. The sentiment scores lie in the range (-3, 3) which

Table 2 Performance of proposed DHF algorithm with *IEMOCAP* dataset

Models	Precision	Recall	F-Measure	7 class Accuracy
MFN Algorithm	*89.20*	*85.70*	*85.60*	86.83
TFN Algorithm	87.30	88.60	89.20	88.37
MRN Algorithm	90.00	89.20	93.70	90.97
MRMF Algorithm	93.00	83.50	95.50	90.67
MV-LSTM Algorithm	92.40	90.60	94.30	92.43
Proposed DHF Algorithm	***96.70***	***96.50***	***95.40***	**96.20**

Table 3 Performance of proposed DHF algorithm with *CMU-MOSI* dataset

Models	Precision	Recall	F-Measure	7 class Accuracy
MFN Algorithm	***87.20***	***83.40***	***82.00***	84.20
TFN Algorithm	88.40	84.30	80.30	84.33
MRN Algorithm	89.30	83.40	87.90	86.87
MRMF Algorithm	85.30	89.30	86.90	87.17
MV-LSTM Algorithm	90.60	88.50	89.70	89.60
Proposed DHF Algorithm	***95.60***	***94.30***	***97.50***	**95.80**

Table 4 Performance of proposed DHF algorithm with *CMU-MOSEI* dataset

Models	Precision	Recall	F-Measure	7 Class Accuracy
MFN Algorithm	***88.40***	***86.50***	***75.40***	83.43
TFN Algorithm	88.90	80.00	79.30	82.73
MRN Algorithm	88.50	82.30	85.60	85.47
MRMF Algorithm	92.30	82.40	95.50	90.07
MV-LSTM Algorithm	89.60	92.30	86.50	89.47
Proposed DHF Algorithm	***98.70***	***95.30***	***92.60***	**95.53**

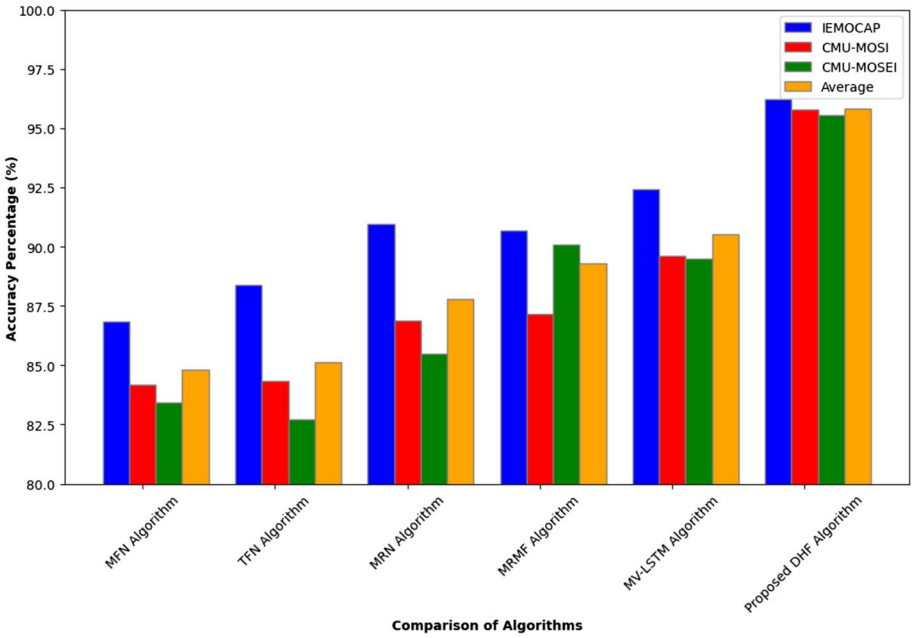

Fig. 5 Accuracy comparison of baseline models against proposed DHF algorithm

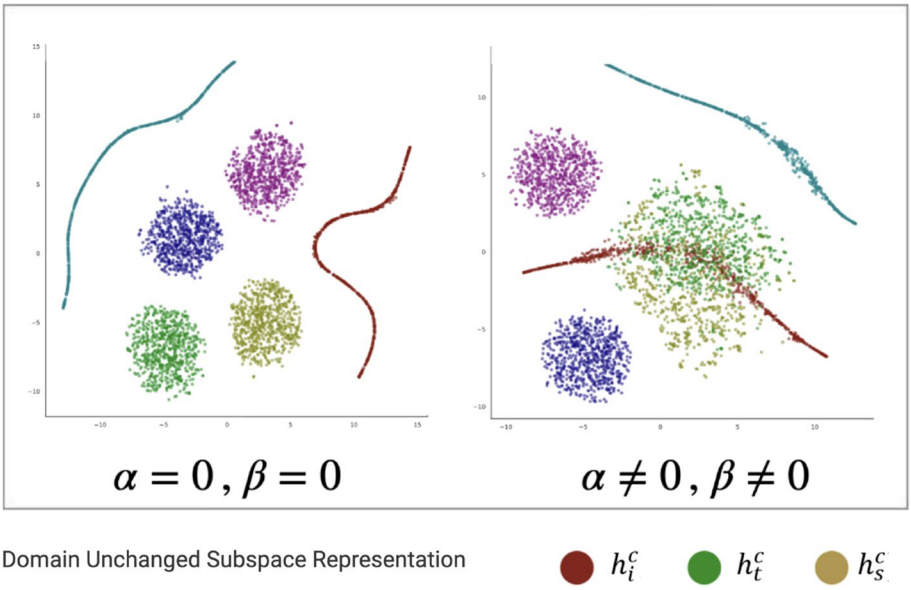

Fig. 6 Visualization of Modality Unchanged Representaion for *CMU-MOSI* dataset

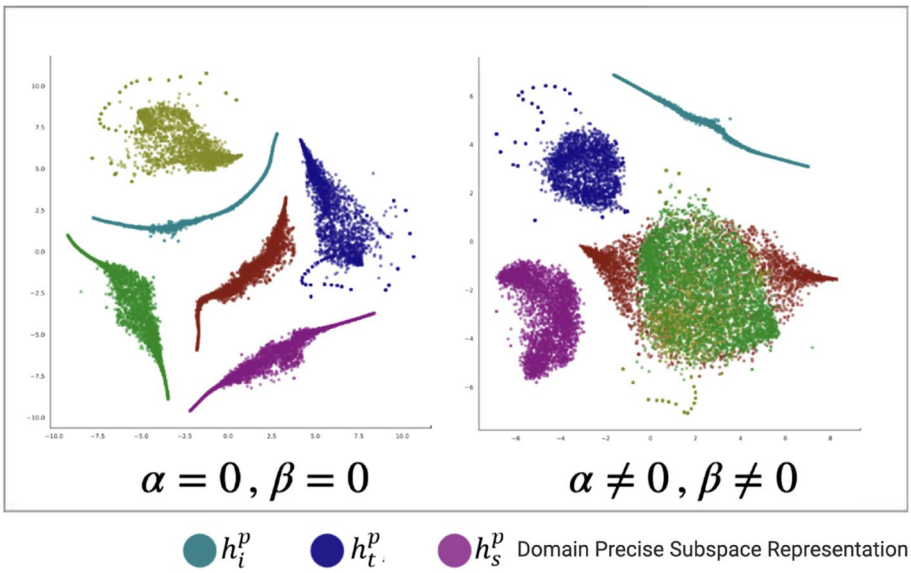

Fig. 7 Visualization of Modality Precise Representation for *CMU-MOSI* dataset

lie between Joyful (3), Surprise (2), Contempt (1), Anger (-3), Disgust (-2), Frustrated (-1) and Neutral (0) sentiments. There are basically three features considered here they are Image, Text and Speech.

A detailed comparative study is performed with proposed algorithm against different baseline model as listed below:

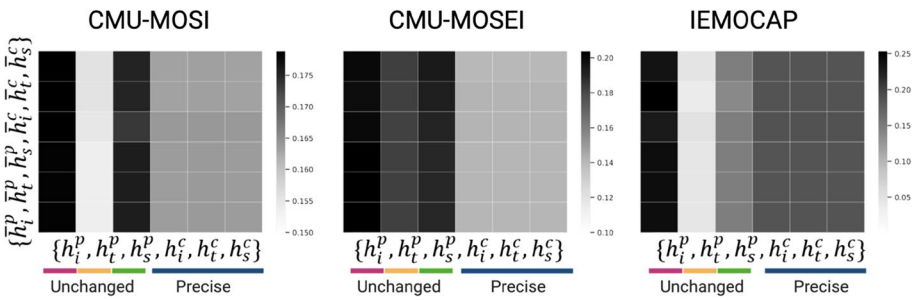

Fig. 8 Visualization of modality unchanged-precise representaion in subspace

- Memory Fusion Network [MFN] and Tensor Fusion Network [TFN] – Used for Temporal and special data analysis.
- Multi-attention Recurrent Network [MRN] and Multimodal Language Analysis with Recurrent Multistage Fusion [MRMF] – Type of Recurrent Neural Network with fusion technique.
- Multi-View Long Short Term Memory [MV-LSTM] – Hierarchical Fusion Analysis algorithm.

The evaluation metrics used for Multimedia Video Analytics with Humour Detection are Precision, Recall, F-Measure and Accuracy. Tables 2, 3 and 4 provide the overview of overall accuracy of baseline models against proposed DHF algorithm by considering IEMOCAP, *CMU-MOSI* and *CMU-MOSEI* Datasets.

Also Fig. 5 represents the overall accuracy of all the baseline models used for analysis against the proposed DHF algorithm which outperforms to provide best accuracy of 95.84%.

The Modality Representation done using proposed "MUR Algorithm" for the benchmark dataset CMU-MOSI as given in Figs. 6 and 7 using tSNE Projections. This figure

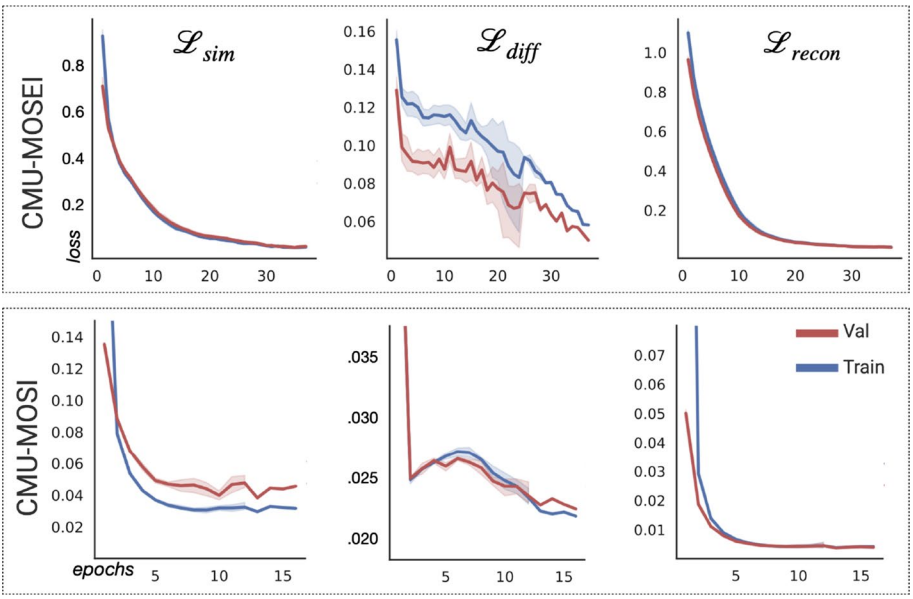

Fig. 9 Loss function for *CMU-MOSI* dataset and *CMU-MOSEI* dataset

depicts all the hidden vectors generated using the feed forward neural network for text, image and speech features.

Figure 8 depicts the representation of six hidden vectors in common subspace as Modality Unchanges-Precise Representaion using MUR Algorithm.

Here the rows of matrix represent Queries and Column Value represents Keys. Hence the input feature vector $\{h_i^p, h_t^p, h_s^p, h_i^c, h_t^c, h_s^c\}$ in the column generate output feature vectors $\{\overline{h}_i^p, \overline{h}_t^p, \overline{h}_s^p, \overline{h}_i^c, \overline{h}_t^c, \overline{h}_s^c\}$.

Figure 9 depicts the trend analysis of Loss function of *CMU-MOSI* and *CMU-MOSEI* Datasets during training. The overall loss in Training (Red line) and Validation.

5 Conclusion and future scope

This research introduced a novel framework for Multimedia Video Analytics, leveraging advanced techniques such as Deep Hybrid fusion Algorithm to significantly enhance the analysis of complex video data. The value of this work was evident in its innovative three-pronged approach, which included Video Feature Extraction, Modality Representation, and Fusion, each contributing uniquely to the accuracy and depth of humor detection. The Enhanced-BERT model was utilized for textual analysis, Multinet-101 for image feature extraction, and the Librosa Library for speech processing. This combination ensured that critical information from text, images, and audio was captured with high precision, addressing the multifaceted nature of humor effectively. A notable contribution was the "Modality Unchanged Precise Representation Learning" strategy within the Modality Representation module. This technique successfully reduced modality gaps and integrated modality-specific characteristics, allowing the system to achieve a more comprehensive understanding of humor. The approach of learning shared and private representations within a subspace enhanced the framework's ability to recognize subtle and context-dependent humor nuances. After representing the modalities in the subspace a novel fusion algorithm called "Deep Hybrid Fusion" is introduced to fuse the hidden vectors represented in the subspace using attention-based transformation method using Softmax suppression. Later to gain knowledge about these subspaces, a mixture of various types of droppings or loss functions are introduced which includes orthogonal loss, reconstruction loss, and distributional similarity loss. In comparison with previous works it is observed that Domain Separation Network learns factorized representations across different instances, whereas proposed model learns the depictions for modalities within inputs. Proposed approach was compared baseline models by considering IEMOCAP, *CMU-MOSI* and *CMU-MOSEI* Datasets which gave best accuracy of 95.85%. Supplementary orthogonal losses across private representations are included by stating that, using both the modality representations aids the fusion by providing a wholesome view of the data. In future, the proposed work can be extended for better modality representations and loss function regularization techniques to improve accuracy.

Acknowledgements This research was supported by M. S. Ramaiah Institute of Technology (MSRIT), Bangalore-560054 and Visvesvaraya Technological University, Jnana Sangama, Belagavi-590018.

Funding The authors did not receive support from any organization for the submitted work.

Data availability The datasets generated during and/or analysed during the current study are available in the MultiComp Lab repository, http://multicomp.cs.cmu.edu/resources/

Declarations

Ethics approval We did not use animals and Human participants in the study reported in this work.

Informed consent For this type of study informed consent is not required.

Consent for publication For this type of study consent for publication is not required.

Competing interest There is no conflict of interest.

References

1. Abdu SA, Yousef AH, Salem A (2021) Multimodal video sentiment analysis using deep learning approaches, a survey. Inf Fusion 76:204–226
2. Agüero-Torales MM, Salas JIA, López-Herrera AG (2021) Deep learning and multilingual sentiment analysis on social media data: an overview. Appl Soft Comput 107:107373
3. Pandian AP (2021) Performance evaluation and comparison using deep learning techniques in sentiment analysis. J Soft Comput Paradigm (JSCP) 3(02):123–134
4. Basiri ME, Nemati S, Abdar M, Asadi S, Acharrya UR (2021) A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. Knowl-Based Syst 228:107242
5. Yafooz WM, Alsaeedi A, Alluhaibi R, Abdel-Hamid ME (2022) Enhancing multi-class web video categorization model using machine and deep learning approaches. Int J Electr Comput Eng 12(3):3176
6. Zheng L, Wang Y, Wang J, Huang Y, Zhang J (2022) Multi-modal semantic video analysis using graph convolutional networks. IEEE Trans Image Process 31:1368–1382. https://doi.org/10.1109/TIP.2021.3125101
7. Li Y, Wang L, Wang H, Sun Y (2022) Deep learning-based multimedia analytics for video content analysis. IEEE Trans Multimed 24(1):1–1. https://doi.org/10.1109/TMM.2021.3113772
8. Kumar S, Singh S, Singh SK (2022) A hybrid approach for multimedia video analytics using semantic segmentation and object detection. Multimed Tools Appl 81(2):677–698. https://doi.org/10.1007/s11042-021-13014-4
9. Chen Y, Li X, Wang Y, Chen C (2022) Spatio-temporal attention network for action recognition in videos. IEEE Trans Neural Netw Learn Syst 33(2):462–472. https://doi.org/10.1109/TNNLS.2021.3082173
10. Zhang S, Huang G, Liu F, Qiao Y (2022) Multi-task learning for object detection and tracking in videos. IEEE Trans Image Process 31:640–652. https://doi.org/10.1109/TIP.2021.3122699
11. Makantasis K, Georgogiannis A, Voulodimos A, Georgoulas I, Doulamis A, Doulamis N (2021) Rank-r fnn: a tensor-based learning model for high-order data classification. IEEE Access 9:58609–58620
12. Liu J, Huang Z, Xu X, Zhang X, Sun S, Li D (2020) Multi-kernel online reinforcement learning for path tracking control of intelligent vehicles. IEEE Trans Syst, Man, Cybern: Systems 51(11):6962–6975
13. Wang Z, Gao P, Chu X (2022) Sentiment analysis from customer-generated online videos on product review using topic modeling and multi-attention BLSTM. Adv Eng Inform 52:101588
14. Ryffel T, Trask A, Dahl M, Wagner B, Mancuso J, Rueckert D, Passerat-Palmbach J (2018) A generic framework for privacy preserving deep learning. *arXiv preprint* arXiv:1811.04017
15. Zhang D, Wei S, Li S, Wu H, Zhu Q, Zhou G (2021) Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 16, pp 14347–14355
16. Yin Y, Meng F, Su J, Zhou C, Yang Z, Zhou J, Luo J (2020) A novel graph-based multi-modal fusion encoder for neural machine translation. *arXiv preprint* arXiv:2007.08742
17. Zhu Y, Xu W, Zhang J, Liu Q, Wu S, Wang L (2021) Deep graph structure learning for robust representations: a survey. *arXiv preprint* arXiv:2103.03036
18. Li D, Rzepka R, Ptaszynski M, Araki K (2020) HEMOS: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media. Inf Process Manage 57(6):102290
19. Wadawadagi R, Pagi V (2020) Sentiment analysis with deep neural networks: comparative study and performance assessment. Artif Intell Rev 53(8):6155–95
20. Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 31(7):1235–1270

21. Behera RK, Jena M, Rath SK, Misra S (2021) Co-LSTM: convolutional LSTM model for sentiment analysis in social big data. Inf Process Manage 58(1):102435
22. Huddar MG, Sannakki SS, Rajpurohit VS (2020) Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification. Comput Intell 36(2):861–881
23. Hu J, Liu Y, Zhao J, Jin Q (2021) MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. arXiv preprint arXiv:2107.06779
24. Dresvyanskiy D, Ryumina E, Kaya H, Markitantov M, Karpov A, Minker W (2020) An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild. *arXiv preprint* arXiv:2010.03692
25. Ranjan B, Sun W, Park J, Mishra K, Schmidt F, Xie R, ... Prabhakar S (2021) DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. Nat Commun 12(1):1–12
26. Han W, Chen H, Gelbukh A, Zadeh A, Morency LP, Poria S (2021) Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In: Proceedings of the 2021 International Conference on Multimodal Interaction pp 6–15
27. Li Z, Li X, Wei Y, Bing L, Zhang Y, Yang Q (2019) Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. *arXiv preprint* arXiv:1910.14192
28. Zhao P, Hou L, Wu O (2020) Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. Knowl-Based Syst 193:105443
29. Shafiei G, Baillet S, Misic B (2022) Human electromagnetic and haemodynamic networks systematically converge in unimodal cortex and diverge in transmodal cortex. PLoS Biol 20(8):e3001735