Artificial Intelligence Review (2024) 57:34
https://doi.org/10.1007/s10462-023-10645-7

An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network

Bairavel Subbaiah[1] · Kanipriya Murugesan[2] · Prabakeran Saravanan[3] · Krishnamurthy Marudhamuthu[4]

Published online: 5 February 2024

Abstract
Sentiment analysis is a key component of many social media analysis projects. Additionally, prior research has concentrated on a single modality in particular, such as text descriptions for visual information. In contrast to standard image databases, social images frequently connect to one another, making sentiment analysis challenging. The majority of methods now in use consider different images individually, rendering them useless for interrelated images. We proposed a hybrid Arithmetic Optimization Algorithm- Hunger Games Search (AOA-HGS)-optimized Ensemble Multi-scale Residual Attention Network (EMRA-Net) technique in this paper to explore the modal correlations including texts, audio, social links, and video for more effective multimodal sentiment analysis. The hybrid AOA-HGS technique learns complementary and comprehensive features. The EMRA-Net uses two segments, including Ensemble Attention CNN (EA-CNN) and Three-scale Residual Attention Convolutional Neural Network (TRA-CNN), to analyze the multimodal sentiments. The loss of spatial domain image texture features can be reduced by adding the Wavelet transform to TRA-CNN. The feature-level fusion technique known as EA-CNN is used to combine visual, audio, and textual information. The proposed method performs significantly better than the existing multimodel sentimental analysis techniques of HALCB, HDF, and MMLatch when evaluated using the Multimodal Emotion Lines Dataset (MELD) and EmoryNLP datasets. Also, even though the size of the training set varies, the proposed method outperformed other techniques in terms of recall, accuracy, F score, and precision and takes less time to compute in both datasets.

Keywords Arithmetic Optimization Algorithm (AOA) · Hunger Games Search (HGS) · Ensemble Multi-scale Residual Attention Network (EMRA-Net) · Multimodal sentiment analysis · Feature extraction

1 Introduction

User-generated content is often posted by users in a variety of formats, contributing to content diversity (Liu et al. 2020). Sentiment analysis is essential for evaluating individual behavior and has many applications, such as review analysis, product analysis, and mental

Extended author information available on the last page of the article

health therapy (Kumar and Garg 2019). Sentiment analysis has grown to be crucial for monitoring people's attitudes and emotions by examining this unstructured, multimodal, informal, high-dimensional, and noisy social data (Xuanyuan et al. 2021). In contrast to traditional social media, such as newspapers, online social media is packed with a plethora of information from many sources that can provide substantially more signs to evaluate sentiments than information provided by words alone. The widespread usage of smartphones in the current world increases the number of users willing to post a multimodal message on social media (Baecchi et al. 2016). Sentiment analysis may be used to determine the polarity of a sentiment, or whether it is positive, negative, or neutral, as well as its emotion, or where it fits on the emotional spectrum type, such as happy or sad, or to define sentiment intensity (Yan et al. 2022).

When it comes to sentimental analysis, the area of Natural Language Processing (NLP) is faced with significant complexities (Tembhurne and Diwan 2021). Machine learning methods are typically helpful for identifying and predicting whether a document shows positive or negative sentiment. The two categories into which machine learning is separated are supervised and unsupervised machine learning algorithms (Tripathy et al. 2015). The social media dataset is simpler to train and comprehend using machine learning. Beyond machine learning, rule-based and lexicon-based techniques are the most often used methodologies. Using a range of classifiers, including deep neural networks (DNN), artificial neural networks (ANN), multilayer perception (MLP), deep neural networks (DNN), and others, the multimodal sentiment analysis approaches place an emphasis on analyzing strong traits individually (Bairavel and Krishnamurthy 2020).

The single modality sentiment classification techniques (Cambria et al. 2017) mainly focus on analyzing the textual or visual content based on its interrelation with its target class and often fail to accommodate the use of more than one feature (acoustic and visual features). Since social media data comprises a diversity of information, sentiment classification utilizing a single modality does not always leads to an optimal sentiment analysis decision. For an effective social media multimodal sentiment analysis, the relationship between the target class, textual content, acoustic feature, and visual features needs to be integrated which is often neglected by most of the existing studies (Cambria et al. 2013; Stappen et al. 2021).

Motivated by these issues, we have planned to efficiently model these interactions in social media posts. Multimodal sentiment analysis mainly relies on more than one type of modality and integrates different such as video, audio, and image modalities for performing sentimental analysis. The difference between single and multimodal analysis is that the extraction of sentiments completely from a single modality is easier when extracting the sentiments from different modalities (Lopes et al. 2021). Since sentiments are easier to extract from text-only data than from data that contains both text and images, this is quite challenging in the context of sentiment analysis. This is a challenging endeavor, particularly when using social media data, as the various data kinds are sparse, might have a variety of contexts, and purposes, and convey irony, and their integrated evaluation is not simple. However, as compared to a text-only strategy, a multimodal approach can enhance performance.

One of the novelties associated with our proposed model is the usage of the Ensemble Attention CNN (EA-CNN) technique for a simple and reliable fusion, The EA-CNN technique investigates the text, acoustic, and visual components of the post individually, and the final classifier is mainly built on the top of these features. The proposed model mainly conducts multimodal sentimental analysis in social media data which is often sparse and has diverse contexts. The state-of-art techniques often fail to utilize the textual, acoustic,

and visual information at once. Multimodal posts in social media have different forms such as audio, video, text, and image. Hence, an end-to-end model that efficiently captures the intra and inter-modality interactions are developed in this paper.

In this paper, we introduce a novel method called Hybrid AOA-HGS optimized Ensemble Multi-scale Residual Attention Network (EMRA-Net), which enhances multi-modal feature fusion by obtaining more dynamic multi-modal information and can more precisely predict emotional intensity. Prior to using multi-modal modeling to extract the context representation of multimodal information, each sentence in the video is independently examined for textual, audio, and visual components. Additionally, to generate a more accurate prediction of emotion, the Hybrid Arithmetic Optimization Algorithm and Hunger Games Search (AOA-HGS) optimized EMRA-Net completely make use of the dynamic information of intra-modal relational and inter-modal interaction. The main contribution of this paper is described below.

- The hybrid AOAHGS-optimized EMRA-Net technique is proposed for predicting the multimodal sentiments in social media based on audio, video, and text.
- The proposed model individually analyzes the sentiments present in each modality (text, acoustic, and visual) by modeling the inter and intra-modal representations which the existing techniques often failed to accomplish.
- For obtaining the best feature set with the highest accuracy, the combination of the arithmetic optimization algorithm and hunger games search is developed. The hyperparameters of EMRA-Net are optimized through AOA-HGS.
- To analyze the multimodal sentiments, the EMRA-Net utilizes 2 segments such as EA-CNN and TRA-CNN. A wavelet transform is introduced in TRA-CNN for reducing the loss of image and texture features that occurs in the spatial domain.
- Evaluating the efficiency of the proposed hybrid AOAHGS optimized EMRA-Net technique through conducting experiments using MELD and EmoryNLP datasets.

The rest of the paper is arranged accordingly. The related works are described in Sect. 2. Section 3 illustrates the proposed methodology. The experimentation results are evaluated in Sect. 4. At last, the conclusions of the paper are described in Sect. 5.

2 Related works

Zhao et al. (2019) developed an image-text consistency-driven multimodal sentiment analysis approach to address the challenge of how efficiently employing information from both visual and text-based postings. After this model explores the link between the image and the text, a multimodal adaptive sentiment analysis approach was applied. A machine-learned sentiment analysis technique was developed by merging textual, visual, and social components with mid-level visual features obtained using the classic sentibank approach to represent visual concepts. Nevertheless, expressing an image's features is a significant challenge. Bairavel et al. (2020) introduced an audio–video-textual-based multimodal sentiment analysis for social media. This model investigates sentiments that were extracted from web recordings using text, audio, and video modalities. In order to combine the retrieved features from several modalities, a feature-level fusion technique is used. The best characteristics from the retrieved data were selected by utilizing an oppositional grass bee optimization (OGBEE) algorithm, finding the best possible feature set. For sentiment

classification, this model used a multilayer perceptron-based neural network but it requires more computational time.

A multi-modality framework called Hierarchical self-attention Fusion-Contextual Self-attention Temporal Convolutional Network (H-SATF-CSAT-TCN-MBM) was developed by (Xiao et al. 2020) for sentiment analysis in the social internet of things. To improve the performance of the CSAT-TCN model in long memory problems, multi-branch memory networks were used, the self-Attention Fusion framework (H-SATF) was used to fuse multi-modality features, and the CSAT-TCN was used to capture the internal and external correlation of multi-modality features. However, this model was unable to learn sentimental qualities in both dual and single modalities at the same time. A Hierarchical Deep Fusion (HDF) model was presented by Xu et al. (2019) to investigate the cross-modal correlations between images, texts, and their social connections. They used Long Short-Term Memory (LSTM) with three levels to find the intermodal links between image and text scaled differently by integrating visual content with multiple textual semantic fragments. A weighted relation network was used to characterize the links between social media images, and each node was embedded in a distributed vector to make the most efficient use of the link information. This model, however, does not include a heterogeneous network embedding mechanism to provide better encapsulation of the network topology.

To analyze the multimodal sentiment, Li et al. (2021) developed a Hierarchical Attention LSTM technique based on the Cognitive Brain limbic system (HALCB). The usage of a Hash algorithm improved the retrieval speed and accuracy. The Random Forest (RF) was trained to recognize and understand the regular distribution of previous outputs before altering the classification results. The three datasets used for testing were YouTube, MOSEI, and MOSI. The HALCB outperformed the other existing techniques in both multi and binary classification tasks. Furthermore, the high path module's sub-network fault tolerance capabilities were not improved. The Deep Multimodal Attentive Fusion (DMAF) method was investigated by Huang et al. (2019) for the analysis of image-text sentiment. The semantic attention approach was used to address the emotion-based words, and the visual attention method was used to treat the emotional areas automatically. The datasets utilized for testing were Flickr-m, Twitter, Flickr-w, and Getty. The findings demonstrated that the DMAF was a useful tactic for handling imperfect multimodal data contents in order to predict attitudes.

Yu et al. (2019) presented a target-dependent social media sentimental analysis method named Entity Sensitive Attention and Fusion Network (ESAFN). It analyzed the sentiments present in videos, images, and user profiles in addition to texts. The LSTM model was utilized to determine the hidden state of each word. The Multimodal fusion layer was then used to merge visual and textual representations after the learning of visual representations. In the end, the softmax function was used to classify sentiment. The two multimodal Named Entity Recognition (NER) datasets used to verify the approach showed that it outperformed other competing multimodel classification methods in terms of performance. Paraskevopoulos et al. (2022) presented a feedback module MMLatch that permits top-down cross-modal modeling of interactions between the architecture's lowest layer and upper layer. To allow the model to perform top-down feature masking for each modality, the MMLatch system acquires high-level representations, which are subsequently used to mask the sensory inputs. The MMLatch model was used to identify multimodal sentiments on the CMU-MOSEI dataset.

Zhang et al. (2021) presented a multimodal emotion recognition model for conversational videos based on reinforcement learning and domain knowledge (ERLDK) by integrating domain learning and reinforcement learning concepts. They mainly identified the

emotions of the samples by analyzing the conversations for a prolonged period at a dialogue level. The dialogues are mainly extracted using a window size of three. The multimodal inputs are mainly the semantic, visual, and audio-based features. The recognition accuracy of the classifier is mainly tested by varying the conversation lengths on the two public datasets. The existing techniques are summarized in Table 1 based on the techniques utilized, modalities focused, the dataset used, benefits, and limitations.

Even though the existing techniques offer improved performance they fail to analyze certain important issues which are discussed as follows. The state-of-art techniques often utilize the textual content present along the audio and visual information to generate the sentiment labels for the images taken for analysis and instead only focus on the visual content. In our proposed model, we give equal importance to all three modalities (text, video, and audio) taken for analysis. The EMRA-Net technique can offer fine-grained analysis and identify the crucial connection that exists between the texture, text, and audio features. Bidirectional Encoder Representation from Transformer (BERT) model (Murfi et al. 2022 and Deng et al. 2022) is a pre-trained model which is mainly trained using domain-specific corpora such as Wikipedia and BookCorpus. The BERT and our proposed TRA-CNN model have many similarities. Both our TRA-CNN and BERT analyze the importance of the semantics of each word in the sentence. Both models obtain the text representations by analyzing the semantics and word position in the sentences. Both architectures use the attention mechanism which computes the importance of each word in the input sequence. Since the CNN architecture alone cannot extract the structural and semantic interrelation between the text, the Three-scale Residual Attention module is integrated with the CNN architecture in the proposed model. The three-scale residual attention module can effectively extract the global semantic features from the text. Whereas in BERT, the self-attention operation offers improved machine translation results. The spatial understanding of the BERT model is improved using the attention mechanism whereas we integrated the wavelet transform in our proposed model to preserve the spatial understanding of the model.

3 Proposed methodology

In this paper, we proposed a hybrid AOAHGS-optimized EMRA-Net technique to predict multimodal sentiments using two datasets such as MELD and EmoryNLP. The architecture of the proposed model is depicted in Fig. 1. Initially, preprocessing is done for the audio, text, and video inputs for obtaining noise-free data. Then, the important features from the text, audio, and video are extracted to analyze the sentiments. To analyze the multimodal sentiments, the EMRA-Net utilizes 2 segments such as EA-CNN and TRA-CNN. The audio feature, visual feature, and textual features are classified using TRA-CNN. Subsequently, EA-CNN is employed for the multimodal sentiments fusion. In this, the hyperparameters of EMRA-Net are optimized using AOAHGS. For obtaining the best feature set with the highest accuracy, the Arithmetic Optimization Algorithm is integrated with Hunger Games Search algorithm. Finally, the polarity of the sentiments (positive, negative, and neutral) is predicted.

3.1 Preprocessing

For effective classification, the texts are preprocessed initially before being fed into the classifiers. Preprocessing is necessary for the following reasons. (1) The text types acquired

Table 1 Summary of existing literature

Reference	Technique presented	Modality focused	Dataset used	Benefits	Limitations
Zhao et al. (2019)	Multimodal adaptive sentiment analysis technique	Textual and visual modality	Visual sentiment ontology benchmark dataset	Adaptively adjusts test image features for improved sentimental analysis	Audio features are not taken for analysis
Bairavel et al. (2020)	OGBEE	Audio, visual, and text	IEMOCAP, CMU-MOSI and YouTube dataset	Improved classification accuracy	Time complexity needs to be reduced
Xiao et al. (2020)	H-SATF-CSAT-TCN-MBM	Audio, visual, and text channels	CMU-MOSI dataset and MSU	Overcomes the long memory issue and improves the sentiment recognition accuracy	The error weights need to be automatically updated to improve performance
Xu et al. (2019)	HDF	Image and text modalities	Twitter and Flickr datasets	Improves the sentiment analysis in social images by integrating the network information	Information loss due to excluding the social link information
Li et al. (2021)	Hierarchical LSTM	Audio, visual, and words	YouTube, MOSEI, and MOSI	Staged structure improves accuracy	Fault tolerance of sub-networks needs to be improved
Huang et al. (2019)	DMAF	Image and text (visual and semantic information)	Flickr-m, Twitter, Flickr-w, and Getty	Improves semantic prediction by identifying the complementary information from the text and visual modality	The generalization capacity is not tested using video and audio data
Yu et al. (2019)	ESAFN	Textual and visual analysis	Twitter multimodal and unimodal sentiment analysis datasets	Offers improved performance with multiple entities	The error rate is high for samples that contain multiple smiley faces
Paraskevopoulos et al. (2022)	MMLatch	Multimodal sentiment analysis (video transcriptions, visual sequences, and audio sequences)	CMU-MOSEI sentiment analysis dataset	Provides multimodal representation without additional tuning	The model interpretability needs to be improved

Table 1 (continued)

Reference	Technique presented	Modality focused	Dataset used	Benefits	Limitations
Zhang et al. (2021)	ERLDK	Visual, audio, and text	EMOCAPand MELD	The results are not biased toward certain emotions	The recognition capability can be improved via optimization

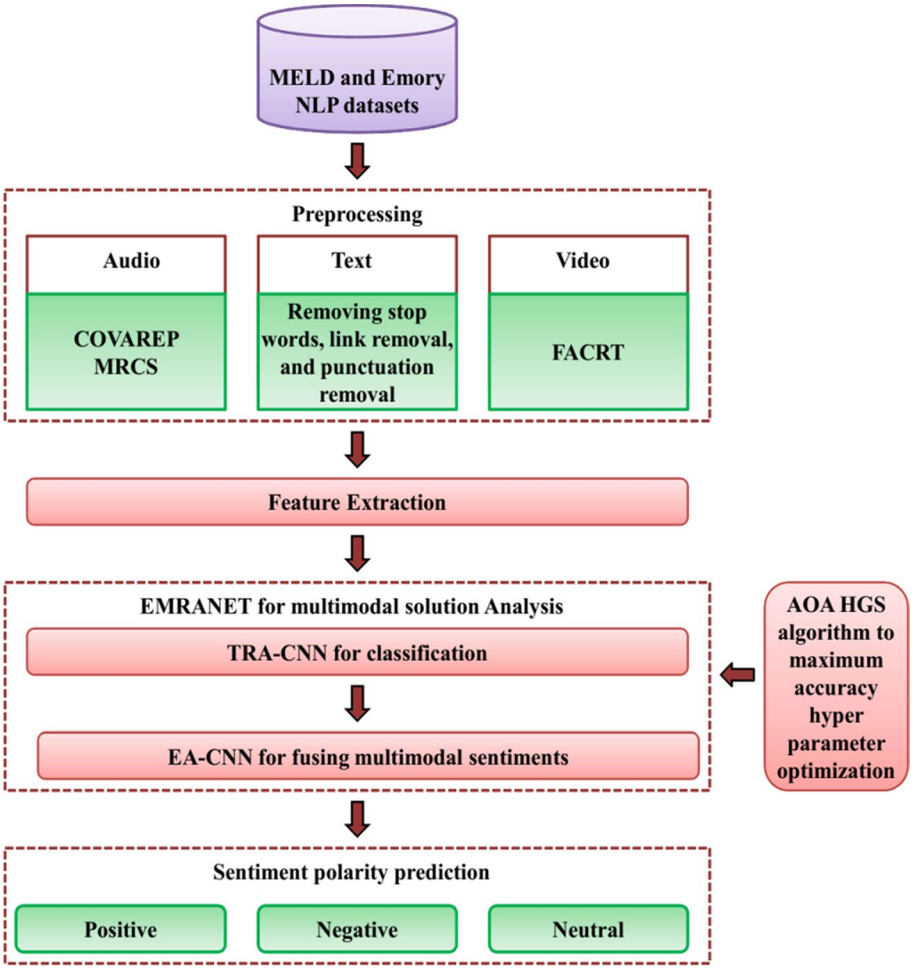


Fig. 1 Overall architecture of proposed hybrid AOAHGS optimized EMRA-Net technique

from social media differ and may include noise, and it contains many semantic and grammatical errors because of their size, slang, and typing speed. Data standardization made it easier for classifiers to learn patterns. (3) Texts will adhere to the input specifications of the Word Embeddings layers and other classifiers. The following are the preprocessing stages: (1) Changing the HyperText Markup Language (HTML) codes as symbols and words, (2) using the Natural Language Toolkit (NTLK) function to remove stop words, (3) changing all words to lowercase, (4) reducing the number of times the same character appears to a maximum of two (for example, changing "Sooohappy" to "so happy"), (5) removing user mentions in social media (for example, the "RT" word on Twitter), and (6) removing punctuation.

The resizing procedure was used to adapt the input images, which had different dimensions, to the standard size of 224 * 224. Subsequently, the mean and standard deviation is applied to normalize the images. It is implemented by subtracting the mean values and subsequently dividing the standard deviation of all images in every channel.

$$I_c = (r_c - \alpha_c)/\sigma_c, \qquad c = 1..,n \tag{1}$$

whereas, r indicates the input image, α is the mean value of the dataset, c refer to the channel, and σ indicates the standard deviation.

3.2 Feature extraction

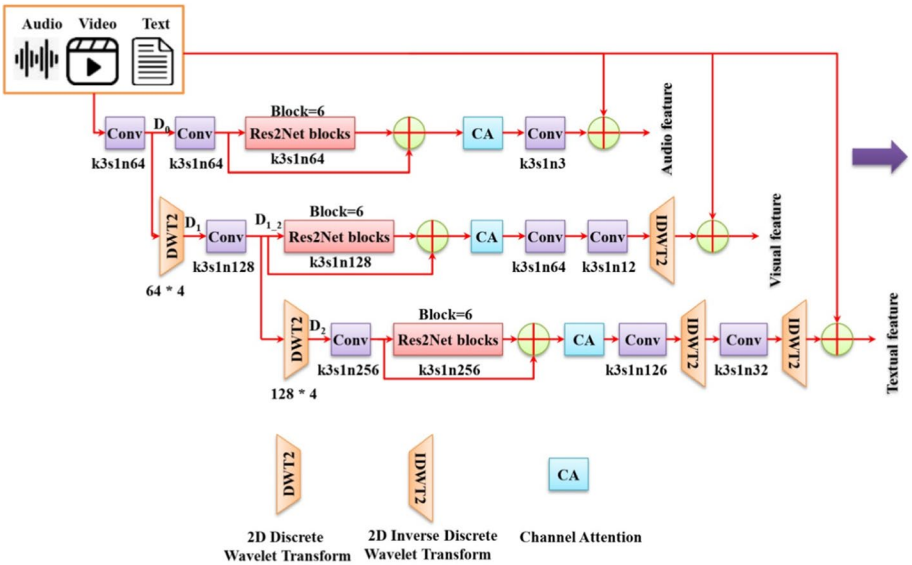
The polarity of the input image is identified by applying the classification and feature extraction over the generated inner representation. We use COVREP for audio to extract low and medium-level acoustic information. Typically, this tool can extract a wide range of rich speech data, including the 12 Mel-Frequency Cepstral Coefficients (MFCCs), maximum dispersion coefficients, peak slope parameters, and voice segment features. First, the user's facial features are extracted from the video utilizing the FACET. A structured feature vector is created by extracting the primary features of the face. Each frame can have it applied to reveal the main facial characteristics while emphasizing and modifying the damaged facial features. Multimodal sentiment analysis proved more significant for numerous tasks including social media analysis. The majority of existing methods, however, only consider the content and are inefficient at capturing the non-linear association across multiple modalities. Despite being vital support for sentiment analysis, connection information between social media images is often neglected, even by those who investigate internal relationships. We concentrate on investigating the multi-modal relationships between text descriptions, visual content, and their social connections for sentiment analysis of social images in order to address these issues.

3.3 EMRA-Net

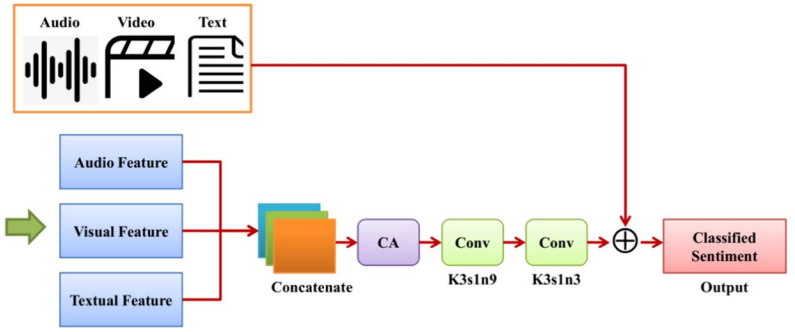
There are 2 segments in the EMRA-Net technique for the analysis of multimodal sentiment using text, audio, and video. They are EA-CNN and TRA-CNN (Wang et al. 2021). The overall structure of EMRA-Net is depicted in Fig. 2. The audio feature, visual feature, and textual features are created using TRA-CNN which is demonstrated in Fig. 2a. All these three multimodal sentiment features are fused through the EA-CNN to predict the sentiments which are depicted in Fig. 2b.

3.3.1 Three-scale residual attention convolutional neural network (TRA-CNN)

3.3.1.1 Three scale branches construction Multi-scale network architecture is more crucial for enhancing recovery image quality. To create various image sizes from input, image pyramiding is a widespread process. However, the loss of image texture features occurs in the spatial domain because of this process. Thus, the upsampling of the low-resolution image into a high-resolution image cannot be reversed. The high-frequency subband information is also maintained by the wavelet transform while performing image downsampling for offering the best solution. In order to create four subband images, four filters are utilized within the 2D DWT (Discrete Wavelet Transform). The wavelet functions and scaling function are multiplied for creating four filters as of 2D DWT's separable property. The three separable wavelet functions and a separable scaling function are described below:



(a) TRA-CNN



(b) EA-CNN

Fig. 2 Structure of EMRA-Net. **a** TRA-CNN, **b** EA-CNN

$$\begin{cases} \gamma(a,b) = \gamma(a)\gamma(b) \\ \chi^{Horiz}(a,b) = \chi(a)\gamma(b) \\ \chi^{Vert}(a,b) = \gamma(a)\chi(b) \\ \chi^{Diag}(a,b) = \chi(a)\chi(b) \end{cases} \tag{2}$$

In the above equation, the one-D wavelet and scaling function is denoted as $\chi(\cdot)$ and $\gamma(\cdot)$. The various 2D wavelet functions are indicated as $\chi^{j=\{Horiz,Vert,Diag\}}(a,b)$, and a 2D scaling function for the low-frequency information is represented as $\gamma(a,b)$. Deviations along diagonals are evaluated by χ^{Diag}(high-frequency information in diagonal), deviations along columns are evaluated by χ^{Horiz} (high-frequency information in horizontal),

and deviations along rows are measured by χ^{Vert}(vertical high-frequency information). The outputs obtained for the input $g(a,b)$ of size $p \times q$ are given as:

$$\begin{cases} R_{\gamma} = \dfrac{1}{\sqrt{p \times q}} \displaystyle\sum_{a=0}^{p-1} \sum_{b=0}^{q-1} g(a,b)\gamma(a,b) \\ R_{\chi}^{t} = \dfrac{1}{\sqrt{p \times q}} \displaystyle\sum_{a=0}^{p-1} \sum_{b=0}^{q-1} g(a,b)\chi^{t}(a,b),\ t = \{Horiz, Vert, Diag\} \end{cases} \tag{3}$$

For the 3 various directions, subband images with high frequency are represented as $R_{\chi}^{t=\{Horiz,Vert,Diag\}}$, and suitable image of the input with low frequency is denoted as R_{γ}. An input image's shallow feature maps are indicated as D_0 in Fig. 2a which is derived using the below equation as,

$$D_0 = \beta(E(H(input))) \tag{4}$$

The batch normalization is indicated as E, the ReLU activation function is denoted as $\beta(\cdot)$, and the 3×3 convolution is denoted as $H(\cdot)$. The downsampled feature map $\left(D_1\right)$ is the input image's half (1/2) scale obtained by decomposing D_0 through the use of the Haar wavelet transform. The redundancy and parameters of an entire network are minimized by using the Convolutional operation behind D_1. The below equation describes D_{1_2} as:

$$D_{1\_2} = \beta\left(E\left(H\left(D_1\right)\right)\right) \tag{5}$$

Finally, feature maps $\left(D_2\right)$ are created for parallel feature learning in TRA-CNN utilizing the three scale branches that are formed from the input image's quarter (1/4) scale, which is likewise derived using the Eq. (5).

3.3.1.2 Deep residual learning module To provide improved sparse learning, the Res2Net modules are deployed within the TRA-CNN. The feature maps of the initial 1×1 convolution are divided into q feature map subsets i.e., $a_j(j \in \{1,2,...,q\})$ by the Res2Net module as illustrated in Fig. 3. Equation (6) defines the output b_j:

$$b_j = \begin{cases} a_j & j = 1 \\ T_j & j = 2 \\ T_j\left(a_j + b_{j-1}\right) & 2 < j \le q \end{cases} \tag{6}$$

The 3×3 convolution layer is indicated as $T_j(\cdot)$. Six Res2Net blocks are implemented in each and every scale branch in TRA-CNN and the four subsets are separated within each module.

3.3.2 Channel attention (CA) block

High-level feature maps for many channels provide fine feature information, and channel information weighting is crucial for multimodal sentiment analysis. The channel attention component is built inside each scale branch of the TRA-CNN in order to maintain the interdependency of the channels underlying the Res2Net blocks. ReLU and Convolutional

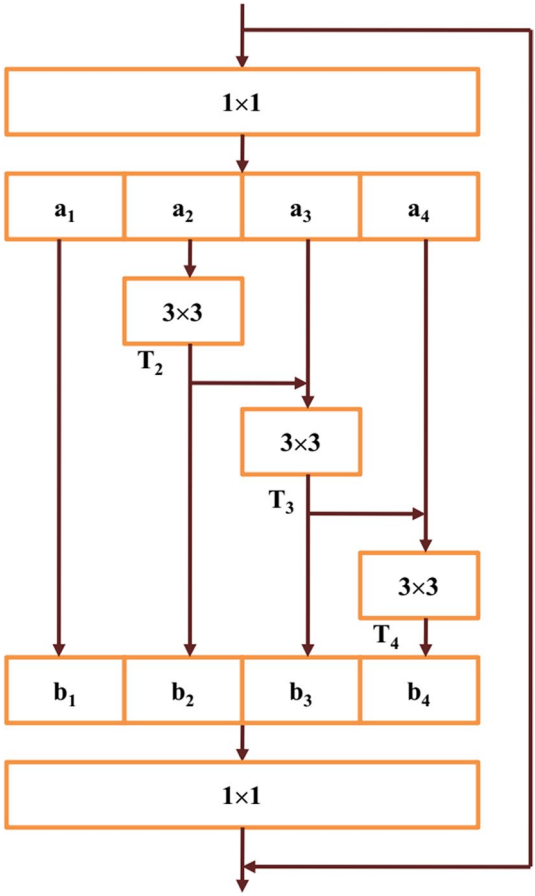


Fig. 3 Res2Net module

operation are used to assess the weights of the different input channels. Finally, element-wise multiplication is used to provide the output for the channel attention component.

3.3.3 Generating multimodal sentiments

The residuals among multimodal input and sentiments present in the input are predicted in every scale branch. After the channel attention part, the skip link and 3×3 convolutions are used in the first scale branch to generate the audio features. Then, the inverse wavelet transforms and convolutional operations are employed in the second and third scale branch for creating the visual and textual features.

3.3.4 Ensemble Attention CNN (EA-CNN)

An efficient multimodal sentiment analysis fusion strategy cannot be achieved by directly integrating features from different scale branches without taking into account different input channel weights. The EA-CNN is created as a result of integrating multimodal sentiments. The channel attention block, which is seen in Fig. 2b, is used to automatically establish the weights for the individual channels once the audio, visual, and textual elements

have first been integrated. Finally, the sentiments contained in the input and the multimodal input still predict the residuals. Due to the lengthy skip link, the EA-CNN and efficient multimodal sentiments fusion also enhance the global sparse learning potential.

3.3.5 Feature level fusion

The advantage of this strategy was that it was rather simple. and generated fundamentally excellent accuracy. The Hybrid AOA-HGS optimized EMRA-Net method combined each methodology's feature vector into a single feature vector stream. Then, each video section is classified into sentiment classes using this vector. Here, the video is used to extract three features: audio, video, and text. Following that, the next models, including audio, video, and text, are extracted, and the resulting modalities are expressed as follows:

$$Audio\,Modality\ :\ Y_{AUDIO} = \left\{Y_1, Y_2, Y_3, \ldots Y_L\right\} \tag{7}$$

$$Vedio\,Modality\ :\ Y_{VIDEO} = \left\{Y_1', Y_2', Y_3', \ldots Y_L'\right\} \tag{8}$$

$$Textual\,Modality\ :\ Y_{TEXT}\left\{Y_1'', Y_2'', Y_3'' \ldots Y_L''\right\} \tag{9}$$

Based on the aforementioned equations, the feature matrix for the audio, video, and textual modalities is represented as Y_{AUDIO}, Y_{VIDEO}, and Y_{TEXT}. Second, three modalities are employed in the feature-level fusion technique's linear combinations of the feature matrix. Assume that G is the new feature matrix.

$$G_1 = \left\{u_1, u_2, u_3, \ldots u_m\right\} \tag{10}$$

$$G_2 = \left\{v_1, v_2, v_3, \ldots v_m\right\} \tag{11}$$

$$G_3 = \left\{w_1, w_2, w_3, \ldots w_m\right\} \tag{12}$$

The fusion matrices for the three modalities are represented by the equation below:

$$G_1 = \tau Y_{AUDIO} + \sigma Y_{VIDEO} \tag{13}$$

$$G_2 = \tau Y_{AUDIO} + \sigma Y_{TEXT} \tag{14}$$

$$G_3 = \tau Y_{VIDEO} + \sigma Y_{TEXT} \tag{15}$$

The feature matrix fusion of the aforementioned equations Including audio, video, and text in all three modal-respectively resented as G_1, G_2 and G_3. Hence, as a result, the following are the mathematical expressions for the values of u_i,v_i, and w_i.

$$u_i = \tau Y_i + \sigma Y_i'' \tag{16}$$

$$v_i = \tau Y_i + \sigma Y_i' \tag{17}$$

$$w_i = \tau Y_i'' + \sigma Y_i'' \tag{18}$$

The values for τ and σ are taken to be $\tau = 1$ and $\sigma = -1$ from the equation above.

3.4 Hybrid AOA and HGS algorithm

Recently, there have been a number of population-dependent strategies developed. Developed solutions are still being tested to address actual issues being utilized in a variety of engineering techniques. Thus, the techniques utilized by researchers require to be significantly changed and enhanced. A more reliable equilibrium that includes optimization and high-quality efficiency is frequently sought after on the basis of significant evolutionary processes. In this study, a hybrid approach is developed by combining the AOA with Hunger Games Search HGS.

3.4.1 Hunger Games Search (HGS) optimization

HGS, a population-dependent optimization technique, has solved limited and unconstrained problems while preserving the features. The subsections describe the various steps in the HGS algorithm.

3.4.1.1 Moving near food The following mathematical formulas were created to simulate the contraction mode and reflect its approaching behavior (Mahajan et al. 2022).

$$\overrightarrow{Y(t+1)} = \begin{cases} \overrightarrow{Y(t)} \cdot (1 + \Re m(1)), & \Re_1 < k \\ \overrightarrow{Z_1} \cdot \overrightarrow{Y_a} + \vec{S} \cdot \overrightarrow{Z_2} \cdot \left| \overrightarrow{Y_a} - \overrightarrow{Y(t)} \right|, & \Re_1 > k, \Re_2 > F \\ \overrightarrow{Z_1} \cdot \overrightarrow{Y_a} - \vec{S} \cdot \overrightarrow{Z_2} \cdot \left| \overrightarrow{Y_a} - \overrightarrow{Y(t)} \right|, & \Re_1 > k, \Re_2 < F \end{cases} \tag{19}$$

The ranges between $-b$, and b is denoted as $\vec{S}$. The random numbers in the interval [0, 1] are represented as $\Re_1$ and $\Re_2$. The current iteration is denoted as t. Random number satisfying normal distribution is denoted by $\Re m(1)$. Hunger's weight is represented by $\overrightarrow{Z_1}$ and $\overrightarrow{Z_2}$. Individuals' entire location is reflected using the variable $\overrightarrow{Y(t)}$ and the starting position is k. The location of a random individual among all the ideal individuals is represented by $\overrightarrow{Y_a}$. The following is the equation for deriving F.

$$F = \text{sech}\left(\left| E(j) - Best_{fitness} \right| \right) \tag{20}$$

here, $j \in 1, 2, ..., m$. Each and every individual's fitness value and the best fitness acquired in the present iteration procedure are represented by $E(j)$ and $Best_{fitness}$. The hyperbolic function $\left(\text{sech}(y) = \frac{2}{e^y + e^{-y}} \right)$ is represented as sech. The equation for $\vec{S}$ is given below:

$$\vec{S} = 2 \times b \times \Re - b \tag{21}$$

$$b = 2 \times \left(1 - \frac{t}{maximum_{iteration}} \right) \tag{22}$$

A random number is represented by the symbol $\Re$ in the range [0, 1]. The largest number in an iteration is symbolized by max $imum_{iteration}$.

3.4.1.2 Hunger role The characteristics of starvation in those who are searching are modeled using mathematical simulations. The equation for $\overrightarrow{Z_1}$ is given below:

$$\overrightarrow{Z_1(j)} = \begin{cases} hungry(j) \cdot \frac{M}{sum_{hungry}} \times \mathfrak{R}_4, & \mathfrak{R}_3 < k \\ 1 & \mathfrak{R}_3 > k \end{cases} \tag{23}$$

The equation for $\overrightarrow{Z_2}$ is given below:

$$\overrightarrow{Z_2(j)} = \left(1 - \text{exponential}\left(-\left|hungry(j) - sum_{humgry}\right|\right)\right) \times \mathfrak{R}_5 \times 2 \tag{24}$$

Each and every individual's hunger is represented using the variable *hungry*. The number of individuals is represented by M. sum_{humgry} is the sum of all of an individual's experiences of hunger. Random numbers between 0 and 1 are represented by $\mathfrak{R}_3$, $\mathfrak{R}_4$ and $\mathfrak{R}_5$. The $hungry(j)$ representation is derived using Eq. (25).

$$hungry(j) = \begin{cases} 0, & OF(j) = Best_{fitness} \\ hungry(j) + hunger_{sensation}, & OF(j) = Best_{fitness} \end{cases} \tag{25}$$

All individual fitness in the current iteration is preserved by $OF(j)$. The equation for $hunger_{sensation}$ is given below:

$$hunger_{threshold} = \frac{E(j) - best_{fitness}}{worst_{fitness} - best_{fitness}} \times \mathfrak{R}_6 \times 2 \times \left(upper_{bound} - lower_{bound}\right) \tag{26}$$

$$hunger_{sensation} = \begin{cases} lower_{bound} \times (1 + \mathfrak{R}), & hunger_{thershold} < lower_{bound} \\ hunger_{thershold}, & hunger_{thershold} \geq lower_{bound} \end{cases} \tag{27}$$

Random numbers between 0 and 1 are represented by $\mathfrak{R}_6$. The hunger threshold is represented by $hunger_{threshold}$. All individual's fitness value is denoted by $E(j)$. The best fitness and worst fitness attained during the current process of iterations are represented by $best_{fitness}$ and $worst_{fitness}$. The search space of the lower bound and the upper bound is represented by $lower_{bound}$ and $upper_{bound}$. There is a lower bound $\left(lower_{bound}\right)$, to the sensation of hunger $\left(hunger_{sensation}\right)$.

3.4.2 Arithmetic optimization algorithm

Basic mathematical operations including division, addition, multiplication, and subtraction are used in a meta-heuristic method known as AOA. To carry out the optimization over numerous search domains, this is both used and modeled. PBAs (population-based algorithms) usually start the process of improving their algorithms by randomly selecting a few candidate techniques. A specific objective function progressively evaluates this stated response while utilizing a set of optimization standards to gradually improve it. The chance of an ideal general solution to the problem is raised by the availability of alternative solutions and optimization simulations. The optimization process is divided into two cycles: exploration and exploitation, taking into consideration variations between meta-heuristic methodologies in PBA approaches.

Along with analysis, geometry, and algebra, arithmetic is one of the most important components of contemporary mathematics. Arithmetic operators (AO) have traditionally

been used in the study of numbers. A few basic mathematical procedures are used while employing optimization to find perfect elements, especially with chosen solutions. The main driving force behind the new AOA is the use of AO to address problems. The optimization procedure starts with a few well-chosen sets, denoted by B in the Eq. (28). In an ideal setting, each iteration is generated at random.

$$B = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & \cdots & b_{1,i} & b_{1,1} & b_{1,m} \\ b_{2,1} & b_{2,2} & \cdots & \cdots & a_{2,i} & \cdots & b_{2,m} \\ b_{2,1} & b_{3,2} & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{M-1,1} & \cdots & \cdots & \cdots & b_{M-1,i} & \cdots & b_{M-1,m} \\ b_{M,1} & \cdots & \cdots & \cdots & a_{M,i} & b_{M,m-1} & b_{M,m} \end{bmatrix} \tag{28}$$

Exploration and exploitation ought to be carefully taken into account at the outset of AOA. The math optimizer's accelerated coefficient is defined by the following equation.

$$MOA\,(D_{iter}) = Min + D_{iter}y\left(\frac{Max - Min}{E_{iter}}\right) \tag{29}$$

where $MOA\,(D_{iter})$ represented as the k^{th} iteration function value,D_{iter} is represented as the current iteration, E_{iter} is represented as a maximum number of iterations, and *Min* and *Max* indicate the accelerated function of Max and Min values.

3.4.3 Exploration stage

The exploratory component of AOA is examined, and it is noticed that, according to AO, calculations using the division or multiplication operators have generated high distribution values or choices that support an exploration search method. These division and multiplication operators never easily attain the aim, in contrast to other operators, because of the high distribution of subtraction and addition operators. Exploiting the search field indiscriminately throughout several areas, AOA exploration operators search for a better choice using the two main search strategies division and multiplication as shown in the equation below.

$$b_{k,i}(D_{iter}+1) = \begin{cases} bestb_i \div (MOP \div \omega) \times ((TV_i - UV_i) \times \lambda + TV_i), p_2 < 0.5 \\ bestb_i \times MOP \times ((TV_i - UV_i) \times \lambda + UV_i), otherwise \end{cases} \tag{30}$$

where, $b_{k,i}(D_{iter}+1)$ represented as i^{th} position in the current iteration, λ represented as a control parameter ≤ 0.5, $b_k(D_{iter}+1)$ that indicates k^{th} solution of the next iteration,ω indicates the smallest integer number, $bestb_i$ represents the k^{th} position of optimum solution attained till now,UV_i *and* TV_i indicates the lower and upper bound limit,

$$MOP(D_{iter}) = 1 - \frac{D_{iter}^{\frac{1}{\beta}}}{S_{iter}^{\frac{1}{\beta}}} \tag{31}$$

where, M_{iter} indicates Max iterations ≤ 5, $MOP(D_{iter})$ represented as k^{th} an iteration function value, Math optimizer Probability (MOP) indicates coefficient. D_{iter} indicates the current iteration. According to AO mathematical formulas, which produced high-density results whether utilizing addition or subtraction, the exploitation nature of AOA is

examined. AOA exploitation operators use two key search approaches to thoroughly scan the field over numerous places in quest of a better alternative. A and S search methods as in the below equation,

Where M_{iter} denotes the maximum number of iterations less than or equal to 5, $MOP(D_{iter})$ is expressed as the value of the k^{th} iteration function, D_{iter} represents the current iteration, and Math Optimizer Probability (MOP) denotes the coefficient. The exploitative character of AOA is looked at in light of AO mathematical formulations which provide high-density results when using addition or subtraction operators. AOA exploitation operators utilize two primary search strategies to thoroughly examine the search space across several locations in pursuit of a better solution. Equation (32) uses the addition and subtraction operators as follows:

$$b_{k,i}(D_{iter}+1) = \begin{cases} bestb_i - MOP \times ((TV_i - UV_i) \times \lambda + UV_i), p_3 < 0.5 \\ bestb_i + MOP \times ((TV_i - UV_i) \times \lambda + UV_i), otherwise \end{cases} \tag{32}$$

3.4.4 Formulation of AOA-HGS

Modern meta-heuristic optimization methods include AOA. AOA is used to address a variety of problems, including those in engineering design, wireless networks, machine learning (ML), power systems, and image processing. The developed strategy is evaluated in light of AOA and HGS. To evaluate performance, each and every strategy is examined using identical parameters, such as the number of iterations and population size. The AOA-HGS technique that has been developed is evaluated by altering the dimensions. A frequent test in earlier research on test function optimization that shows the effect of various dimensions on the efficiency of AOA-HGS is the varied dimension's influence test. This suggests that it is efficient for both high- and low-dimensional issues. In populations of issues with high dimensions, dependent techniques give efficient search results. The implementation of the AOA-HGS model is shown in Fig. 4. From Fig. 4, many phases are used in the established procedure in accordance with the implementation are discussed. Thus, the first phase is defining the various parameters that will be utilized. The solution is produced in the second phase using the specified parameters. Estimating the fitness function is the third phase, and the best solution is selected in the fourth phase. The use of HGS is prohibited in the fifth phase if the random number ($\Re$) value is larger than 0.5, whereas AOA is required if the value is less than 0.5. If the requirements are satisfied in the sixth phase, the optimal solution is sent back to the third phase to calculate the fitness function, and it is then returned in the seventh phase. The complexity of the developed AOAHGS, which is based on the complexity of the original AOA and HGS, is as follows:

$$O(AOA\_HGS) = (M) \times O(AOA) \times O(HGS) \tag{33}$$

$$O(AOA) = O(M \times (t \times \text{dimension} + 1)) \tag{34}$$

$$O(HGS) = O(M \times (t \times \text{dimension} + 1)) \tag{35}$$

The developed AOAHGS's overall complexity is shown below:

$$O(AOA\_HGS) = O(t \times M \times (\text{dimension} + \text{M})) \tag{36}$$

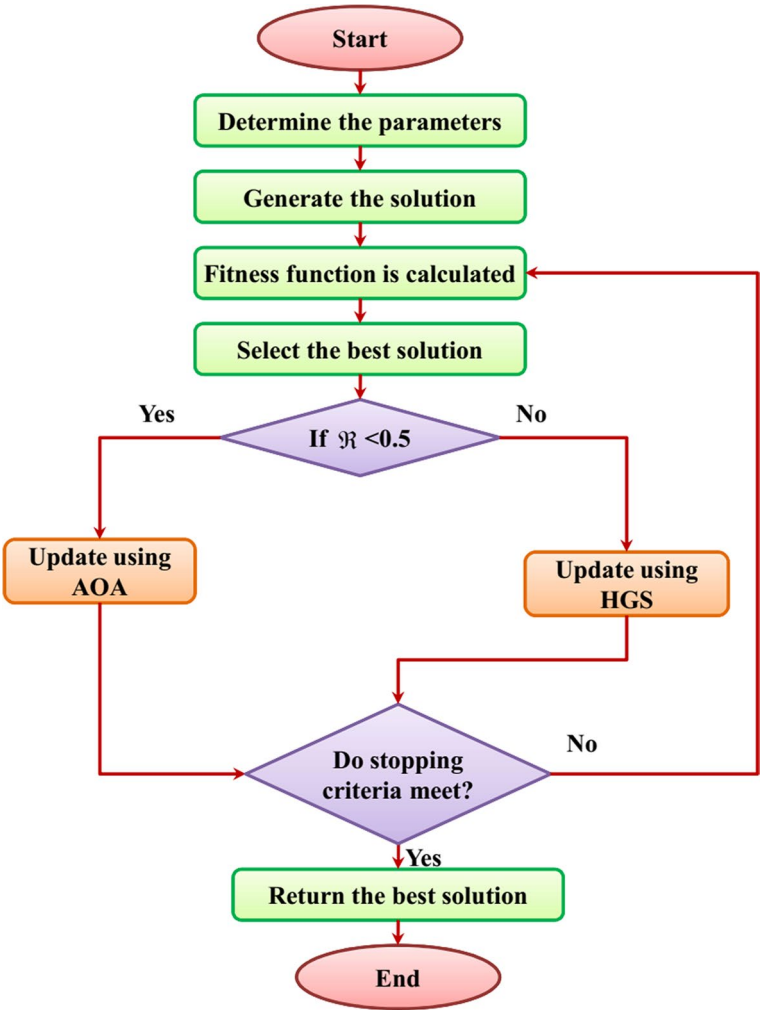


Fig. 4 Implementation of the Hybrid AOA-HGS model

The number of solutions is represented by M. The solution size is denoted by dimension. The number of iterations is represented by t.

3.5 TRA-CNN structure optimization using hybrid AOA-HGS algorithm

The fusion method aims to utilize distinct classification methods for performing the final classification of the images, audio, and text. Thus utilizing the context knowledge from both sources, the fusion strategy seeks to surpass individual classifications. We employed the Hybrid AOA-HGS algorithm to achieve our goal of generating an optimum solution for eliminating the manual processing of the input dataset. Initially, the machine learning model M is defined by a mapping of the architectures Ar and datasets space I to the space models S. The mapping process is depicted as $M : I \times Ar \rightarrow S$ for all datasets, $i \in I$, and

$ar \in Ar$. The mapping reduces the loss function L_f with the associated model s, including architecture ar, parameters ρ, and training data T_d, using the regularisation technique R_t.

$$M(ar, T_d) = \underset{s^{(ar,\rho)} \in S^{(ar)}}{\arg\min} L_f(s^{(ar,\rho)}, T_d) + R_t(\rho) \tag{37}$$

As a result, our method's problem is classified as a nested optimization problem and is solved using the Hybrid AOA-HGS algorithm. It generates an optimal model for classifying the sentiments using the individual classification i and the search space Ar:$ar \ast \in Ar$. It also increases the validation set's objective function ∂.

$$ar \ast = \underset{ar \in Ar}{\arg\max}\, \partial(M(ar, T_d), T_v) \tag{38}$$

Because our task is based on merging text, audio, and image classifications and producing a single output, the first step is to obtain gain I using X_{txt},X_{audio}, and X_{img}. Whereas X_{txt} ,X_{audio}, and X_{img} are the classification outcome for each individual modality. The above three classifiers are integrated as follows.$Y = X_{img} \oplus X_{txt} \oplus X_{audio}$; Whereas Y is the optimization problem input (final classifier). For the three-class sentiment classifications task, ∂ denotes the accuracy.

4 Experimental results and analysis

Experimentation is carried out and reviewed in order to extract audio, video, and textual elements. The proposed Hybrid AOA-HGS optimized EMRA-Net technique methodology is developed in the MATLAB simulation environment (version 2017a). Furthermore, the testing is carried out on a Microsoft Windows 7 Professional computer that is powered by an Intel (R) Core i5 processor with a memory of 16 GB RAM and a clock speed of 3.20 GHz. Accuracy, recall, F-score, and precision are all performance measures. The underlying emotional analysis methods OGBEE (Bairavel et al. 2020), HDF (Xu et al. 2019), H-SATF-CSAT-TCN-MBM (Xiao et al. 2020), HALCB (Li et al. 2021), DMAF (Huang et al. 2019), ESAFN (Yu et al. 2019), and ERLDK (Zhang et al. 2021) are chosen for assessing the proposed method's improvements. The experiments are conducted using different population sizes and iteration values. When the population and number of iterations are set too high such as 200 and 2000, then the time consumption of the proposed model increases with a slight decline in accuracy. Even though the increase in the number of iterations improves the accuracy of the proposed model after some time it lowers. Based on the experimental results, the values set for the hunger threshold, control parameter, and sensitive parameters of the HGS and AOA algorithm are set as 100, 0.5, and 1.5 respectively.

4.1 Dataset description

The different datasets utilized in the study are presented below.

4.1.1 Multimodal emotion lines dataset (MELD)

The MELD was established in order to increase and expand the emotion lines dataset (Ghosal et al. 2019). MELD, like EmotionLines, allows audio and visual content in addition to text and has the same dialogue situations. The MELD dataset (https://github.com/declare-lab/MELD/tree/master/data) comprises about 1400 lines of dialogue and 13,000 utterances from a friend's TV show. Several speakers participated in the discussions. Each phrase in the exchange has been assigned one of these seven emotions: grief, fear, rage, contempt, surprise, neutral, or joy. For each utterance, the MELD contains sentiment annotations such as negative, neutral, and positive.

4.1.2 EmoryNLP dataset

The EmoryNLP dataset is based on the popular television show Friends (Zahiri and Choi 2018). EmoryNLP has 897 scenes, 12,606 utterances, and 97 episodes. Each phrase is tagged with one of the seven emotions drawn from Willcox's (1982) feeling wheel, which contains six fundamental emotions and a default mood of neutral: mad, serene, frightened, sad, powerful, and joyous.

4.2 Performance metrics

Accuracy, precision, recall, and F-score are performance measures that are evaluated, and the computation is as follows.

$$Acc = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \tag{39}$$

$$\text{Re}call = \frac{T_P}{T_P + F_N} \tag{40}$$

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{41}$$

$$F_1 = \frac{2 * \text{Re}call * \text{Pr}ecision}{\text{Re}call + \text{Pr}ecision} \tag{42}$$

4.3 Results analysis

Tables 2, 3, and 4 provide the prediction analysis for three modalities: text, video, and audio. We analyze the accuracy, precision, recall, and F-measure for the three modalities based on positive, negative, and neutral emotions. According to the results of the evaluation, our proposed approach outperforms other current strategies.

Figure 5 compares and analyses the accuracy depending on the text, video, and audio modalities. The proposed approach is contrasted with the three currently used approaches, HALCB, HDF, and MMLatch. When compared to existing multi-model sentiment

Table 2 Prediction performance for text

	Accuracy	precision	F-measure	Recall
Positive	96.63	91.12	90.02	92.77
Negative	96.33	91.43	90.66	92.65
Neutral	96.45	91.07	90.89	92.45

Table 3 Prediction performance for audio

	Accuracy	precision	F-measure	Recall
Positive	96.03	91.92	90.32	92.97
Negative	97.03	91.53	90.76	92.55
Neutral	96.85	91.97	90.69	92.55

Table 4 Prediction performance for video

	Accuracy	precision	F-measure	Recall
Positive	96.93	91.32	90.92	92.97
Negative	96.83	91.73	90.96	92.65
Neutral	96.85	91.67	90.89	92.95

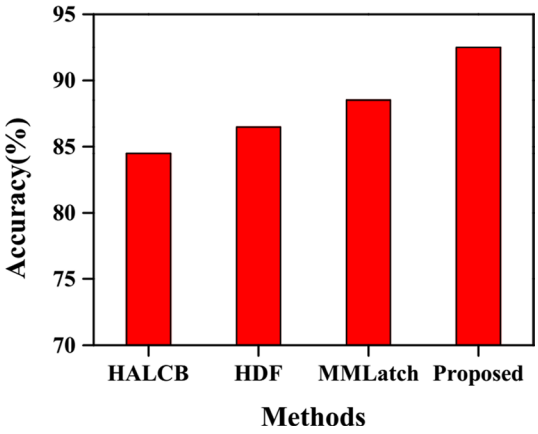


Fig. 5 Analysis of accuracies of various methods

classification models, it demonstrates that the proposed approach delivers a higher level of classification accuracy. The proposed approach has a 94.5% accuracy rate as determined by the figure.

The computational time of all approaches is compared in Fig. 6. The figure reflects that the proposed Hybrid AOA-HGS optimized EMRA-Net approach offers a lower computational time of 1456 s which is lower than the existing three methods. MMLatch takes a computational time of 2128 s, HDF takes 3254 s, and HALCB takes 4255 s.

Table 5 compares the proposed and existing approaches in terms of MELD and EmoryNLP datasets. The three existing algorithms, HALCB, MMLatch, and HDF, are evaluated using these two datasets. Then, our proposed approach is tested against these two

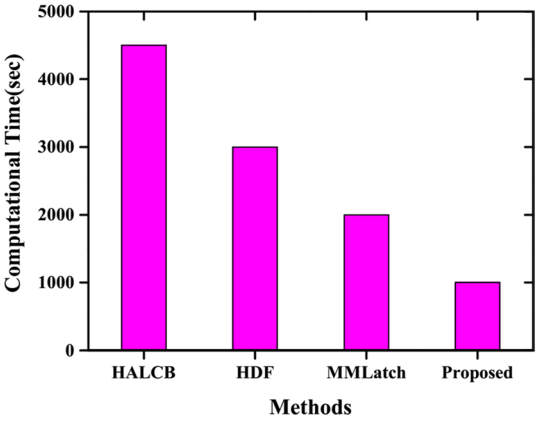


Fig. 6 Analysis of computational time of various methods

Table 5 Comparison of existing technique and proposed method with the dataset

Existing technique	MELD	EmoryNLP dataset
HALCB	75.12	78.56
HDF	85.67	83.41
MMLatch	87.65	87.12
proposed	95.87	96.97

datasets. As a consequence of this comparison, our proposed method outperforms the other methods.

For analyzing the performance deeply we again compare the accuracy of the proposed Hybrid AOA-HGS optimized EMRA-Net method and other existing methods with varying training sizes. The outputs of the experiments conducted on the EmoryNLP Dataset are given in Fig. 7. Based on the graph, we may conclude that the proposed method outperforms the previous three baseline models. It demonstrates that the approach is robust enough for multi-model sentimental analysis.

Figure 8 compares the accuracy of the Hybrid AOA-HGS optimized EMRA-Net approach to the other existing multimodel classification models in the MELD dataset. When the training size is small, all of the techniques depicted in the graph have the same accuracy values. This is due to the fact that smaller data sets cannot be used to effectively train the algorithm. As the magnitude of the training data rises, so do the variations in the accuracy. It demonstrates how the proposed system outperforms other existing models in terms of accuracy. This describes why the proposed method is useful for multimodal sentiment analysis.

The comparison of prediction accuracy utilizing the optimal feature selection approach is shown in Fig. 9. The efficiency of the proposed model in multimodal sentiment analysis is analyzed by comparing it with two existing techniques such as HALCB and HDF. The visual, semantic, and audio modalities are taken for analysis. Based on the investigation, the graph below clearly reveals that our proposed solution is more efficient and performs better.

The ablation study is mainly conducted to test the effectiveness of different components of the proposed model such as Hybrid AOA-HGS, EA-CNN, HGS algorithm, and

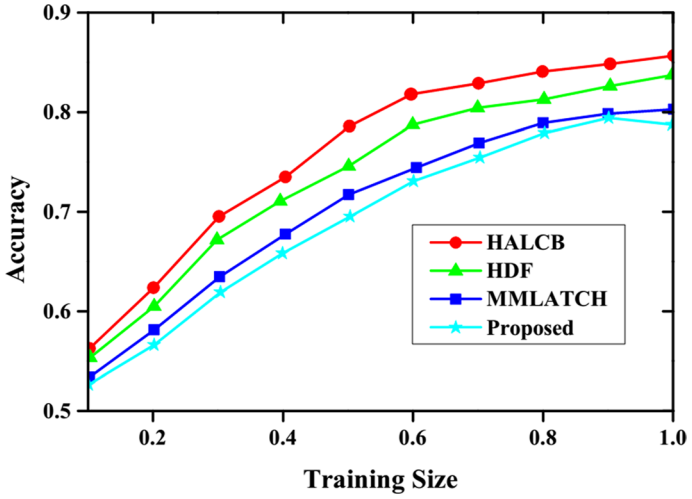


Fig. 7 Analysis of accuracy on varying size training data in of EmoryNLP Dataset

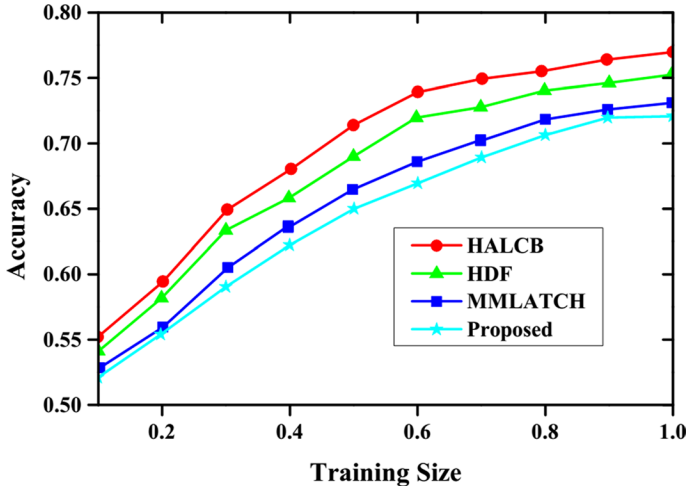


Fig. 8 Analysis of accuracy on varying size training data in of MELDataset

TRA-CNN. The arithmetic mean value in terms of the F1-score computed for each class is termed a macro-averaged F1-score and the accuracy taken is the standard classification accuracy. As per the results shown in Table 6, the EA-CRNN plays a main role in the prediction of the final outcome. Hence, discarding the EA-CNN part will affect the outcome of the multimodal sentimental analysis result and reduces the accuracy to nearly equal to 3% in the MELD and EmoryNLP datasets. Since the TRA-CNN prevents the spatial domain feature loss in images when it is removed from the proposed approach, it results in a decline of accuracy nearly equal to 5% in the MELD dataset and 6% in the EmoryNLP dataset. The effect of the HGS algorithm and the hybrid AOA-HGS algorithm can be

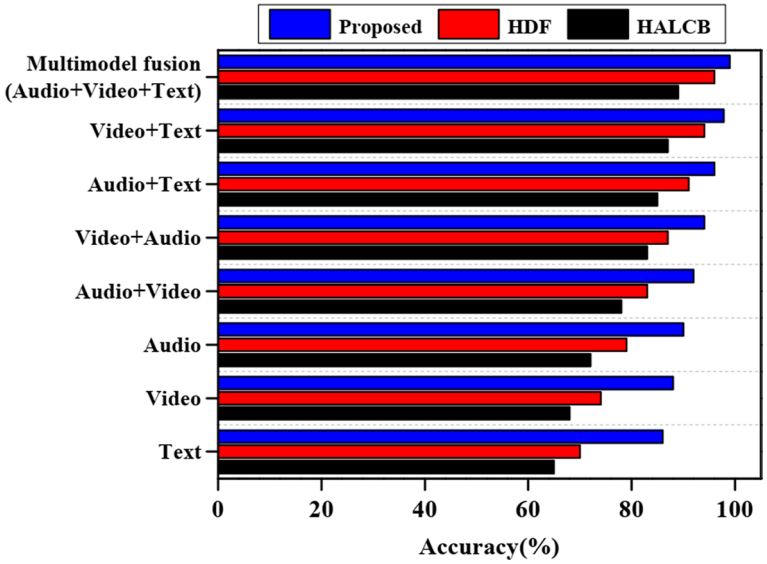


Fig. 9 Comparision of accuracies in various features

Table 6 Ablation study results

Technique	Performance evaluation			
	MELD		EmoryNLP	
	Accuracy	Macro-F1 score	Accuracy	Macro-F1 score
Proposed EMRA-NET	95.87	94.32	96.97	94.65
Without EA-CNN (Using TRA-CNN for classification)	92.56	91.65	91.47	89.63
Without TRA-CNN	90.14	89.54	92.65	89.87
Without HGS algorithm	89.65	88.63	85.45	83.14
Without Hybrid AOA-HGS algorithm	88.45	87.14	87.38	85.23

noticed in the last two columns of Table 6. As per the results, we can see that the removal of the HGS and the Hybrid AOA-HGS algorithm significantly affects the performance with a decline of accuracy from 95.87% to 88.54% in the MELD dataset and 94.65% to 85.23% in the EmoryNLP dataset.

The proposed model is compared with different baseline models such as OGBEE (Bairavel et al. 2020), HDF (Xu et al. 2019), H-SATF-CSAT-TCN-MBM (Xiao et al. 2020), HALCB (Li et al. 2021), DMAF (Huang et al. 2019), ESAFN (Yu et al. 2019), and ERLDK (Zhang et al. 2021). Based on the results demonstrated in Table 7 we can analyze that the proposed model performs well when compared to the baseline models on the MELD dataset. The results demonstrate the efficiency of integrating the visual, audio, and semantic features for multimodal sentiment analysis. The proposed EMRA-NET offers optimal performance due to the usage of the EA-CNN and TRA-CNN techniques which automatically extract the multimodal features which are necessary for identifying accurate

Table 7 Comparative analysis using different performance evaluation metrics

Techniques	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
OGBEE (Bairavel et al. 2020)	89	90	89	87
HDF (Xu et al. 2019)	85	87	88	82
H-SATF-CSAT-TCN-MBM (Xiao et al. 2020)	87	89	90	91
HALCB (Li et al. 2021)	85	87	88	86
DMAF (Huang et al. 2019)	84	86	87	85
ESAFN (Yu et al. 2019)	86	87	88	87
ERLDK (Zhang et al. 2021)	88	89	87	86
Proposed Hybrid AOA-HGS model	93	94	95	93

sentiments. The usage of the hybrid AOA-HGS algorithm shows its efficiency in extracting complementary information from diverse modalities. The state-of-art techniques such as HDF (Xu et al. 2019) and DMAF (Huang et al. 2019) offer optimal performance when evaluated using the integrated textual and visual features when compared to the integration of audio, textual, and visual features. The techniques such as HDF and HALCB offer lower performance when there is a lack of sufficient data.

5 Conclusion

In this study, we presented a new multimodel sentimental analysis named a Hybrid AOA-HGS optimized EMRA-Net for analyzing the sentiments of audio, video, and text inputs. Initially, the proposed method's face characteristics are extracted from each frame for each video segment. The key aspects are then extracted from the textual and visual data for sentiment analysis. The AOA- HGS are then used to optimize the evaluation of the retrieved characteristics. When evaluated using the MELD and EmoryNLP datasets, the proposed model offers higher accuracy for multimodal sentimental analysis techniques when compared to the existing techniques such as HALCB, HDF, and MMLatch. The performance of the proposed model is also higher when evaluated using different performance evaluation measures such as f-score, precision, recall, and accuracy on the two datasets. The computational time computed is also low even with an increase in samples in the training set.

Author contributions KM agreed on the content of the study. BS, KM, PS and KM collected all the data for analysis. KM agreed on the methodology. BS, KM, PS and KM completed the analysis based on agreed steps. Results and conclusions are discussed and written together. The author read and approved the final manuscript.

Funding Not applicable.

Data availability Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Informed consent Informed consent was obtained from all individual participants included in the study.

Consent to participate Not applicable.

Consent for publication Not applicable.

Research involving human and animal participants This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Baecchi C, Uricchio T, Bertini M, Del Bimbo A (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tools Appl 75(5):2507–2525

Bairavel S, Krishnamurthy M (2020) Novel OGBEE-based feature selection and feature-level fusion with MLP neural network for social media multimodal sentiment analysis. Soft Comput 24(24):18431–18445

Cambria E, Das D, Bandyopadhyay S, Feraco A (2017) Affective computing and sentiment analysis. A practical guide to sentiment analysis. Springer, Cham, pp 1–10

Cambria E, Howard N, Hsu J, Hussain A (2013) Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics. In 2013 IEEE symposium on computational intelligence for human-like intelligence (CIHLI). IEEE, pp 108–117

Deng L, Ge Q, Zhang J, Li Z, Yu Z, Yin T, Zhu H (2022) News text classification method based on the GRU_CNN model. International Transactions on Electrical Energy Systems

Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A (2019) Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. Preprint at arXiv:1908.11540

Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019) Image–text sentiment analysis via deep multimodal attentive fusion. Knowl-Based Syst 167:26–37

Kumar A, Garg G (2019) Sentiment analysis of multimodal twitter data. Multimed Tools Appl 78(17):24103–24119

Li Y, Zhang K, Wang J, Gao X (2021) A cognitive brain model for multimodal sentiment analysis based on attention neural networks. Neurocomputing 430:159–173

Liu B, Tang S, Sun X, Chen Q, Cao J, Luo J, Zhao S (2020) Context-aware social media user sentiment analysis. Tsinghua Sci Technol 25(4):528–541

Lopes V, Gaspar A, Alexandre LA, Cordeiro J (2021) An AutoML-based approach to multimodal image sentiment analysis. In 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 1–9

Mahajan S, Abualigah L, Pandit AK (2022) Hybrid arithmetic optimization algorithm with hunger games search for global optimization. Multimedia Tools and Applications, pp 1–24

Murfi H, Gowandi T, Ardaneswari G, Nurrohmah S (2022) BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis. Preprint at arXiv:2211.05273

Paraskevopoulos G, Georgiou E, Potamianos A (2022) Mmlatch: bottom–up top–down fusion for multimodal sentiment analysis. In ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 4573–4577

Stappen L, Schumann L, Sertolli B, Baird A, Weigell B, Cambria E, Schuller BW (2021) Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox. In Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, pp 75–82

Tembhurne JV, Diwan T (2021) Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. Multimed Tools Appl 80(5):6871–6910

Tripathy A, Agrawal A, Rath SK (2015) Classification of sentimental reviews using machine learning techniques. Procedia Comput Sci 57:821–829

Wang J, Li C, Xu S (2021) An ensemble multi-scale residual attention network (EMRA-net) for image Dehazing. Multimed Tools Appl 80(19):29299–29319

Xiao G, Tu G, Zheng L, Zhou T, Li X, Ahmed SH, Jiang D (2020) Multimodality sentiment analysis in social internet of things based on hierarchical attentions and CSAT-TCN with MBM network. IEEE Internet Things J 8(16):12748–12757

Xu J, Huang F, Zhang X, Wang S, Li C, Li Z, He Y (2019) Sentiment analysis of social images via hierarchical deep fusion of content and links. Appl Soft Comput 80:387–399

Xuanyuan M, Xiao L, Duan M (2021) Sentiment classification algorithm based on multi-modal social media text information. IEEE Access 9:33410–33418

Yan X, Xue H, Jiang S, Liu Z (2022) Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling. Appl Artif Intell 36(1):2000688

Yu J, Jiang J, Xia R (2019) Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE/ACM Trans Audio, Speech, Lang Process 28:429–439

Zahiri SM, Choi JD (2018) Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In Workshops at the thirty-second aaai conference on artificial intelligence

Zhang K, Li Y, Wang J, Cambria E, Li X (2021) Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Trans Circuits Syst Video Technol 32(3):1034–1047

Zhao Z, Zhu H, Xue Z, Liu Z, Tian J, Chua MCH, Liu M (2019) An image-text consistency driven multimodal sentiment analysis approach for social media. Inf Process Manag 56(6):102097

Authors and Affiliations

Bairavel Subbaiah[1] · Kanipriya Murugesan[2] · Prabakeran Saravanan[3] · Krishnamurthy Marudhamuthu[4]

✉ Kanipriya Murugesan
mkanipriya@gmail.com

[1] Department of Computer Science and Engineering, KCG College of Technology, Karapakkam, Tamilnadu, India

[2] Department of Computational Intelligence, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, India

[3] Department of Networking and Communications, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, India

[4] Department of Computer Science and Engineering, KCG College of Technology, Karapakkam, Tamilnadu, India