Received: April 7, 2022. Revised: June 28, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

MALO-LSTM: Multimodal Sentiment Analysis Using Modified Ant Lion Optimization with Long Short Term Memory Network

Sri Raman Kothuri[1]* **N R RajaLakshmi[1]**

[1]*Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Avadi-600062, India*
* Corresponding author's Email: sriramankothuri@gmail.com

Abstract: In recent times, multimodal sentiment analysis is the most researched topic, due to the availability of huge amount of multimodal content. Generally, multimodal sentiment analysis uses text, audio and visual representations for effective sentiment recognition. The detection of sentiment in the natural language is a tricky process even for humans, so making it automation is more complicated. In this article, the input multimodal data is collected from Surrey Audio-Visual Expressed Emotion (SAVEE) and YouTube datasets, and then hybrid feature extraction is performed to extract feature vectors from the different modalities like textual, audio and visual. The Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) techniques are applied to extract features from textual modality. Further, AlexNet, spectral centroid features, spectral flux features, and short term energy are employed to extract features from the visual and audio modalities. The extracted active feature values are given as the input to the Modified Ant Lion Optimizer (MALO) based Long Short Term Memory (LSTM) network for sentiment classification. In the MALO algorithm, two new processes are performed to select optimal hyper-parameters of LSTM network that improves the training and testing mechanism and reduces the computational complexity. The results with the MALO-LSTM model obtained were 98.62% and 98.81% of accuracy on YouTube and SAVEE datasets, relatively higher than some of the existing methods, the proposed approach provided comparatively better performance.

Keywords: AlexNet, Computer vision, Long short term memory network, Modified ant lion optimizer, Multimodal sentiment analysis.

1. Introduction

In past few decades, the multimodal sentiment recognition is a growing research topic that plays a vital role in the research fields: computer vision, natural language processing, speech processing, and medical field [1, 2]. Usually, the sentiment detection aims in identifying the attitude or opinion of a speaker or writer towards a general topic such as movie review, product review, political, etc. [3, 4]. Generally, the sentiment of an individual is determined by their facial expression, the spoken words, and emotional tone [5]. Therefore, it is essential in combining visual, acoustic and language modalities for effective sentiment recognition [6, 7]. The existing research studies in sentimental analysis mainly concentrated on holistic-video-fusion that are accomplished with simpler features (for example: bag of words, parts of speech, micro-blogs, etc.) to extract the feature vectors from the different modalities: textual, audio, and visual [8 - 10]. The simple feature fusion methods ignore the textual and audio structures by focusing only on the visual modalities that leads to poor sentiment recognition [11]. For example: if the speakers frequently switch between the topics and opinions in high volatile and tempo videos. In this case, it is hard to understand the opinions convey by the speakers. Correspondingly, an external method is needed to detect the video polarity in the subtlety opinion videos, and to calculate the expressed sentiment's strength [12]. To address these issues, an attempt has been done in multimodal sentiment analysis using deep learning techniques to classify the sentiment of the speakers and readers. Based on the individual's sentiment, an Indian art music raga (Carnatic music) is recommended. The objectives of this study are listed as follows:

- The multimodal input data is acquired from YouTube and SAVEE datasets for an effective sentiment analysis.
- Then, a hybrid feature extraction is performed for extracting feature vectors from the text, audio and visual modalities. The undertaken methods are:

  visual: AlexNet,

  audio: spectral centroid features, spectral flux features and short-term energy, and

  textual: LDA and LSA. The extracted feature vectors reduce the over fitting risk, speed up the training and testing process in the classification model.
- Next, sentiment classification is accomplished using MALO-LSTM model which classifies three sentiments (negative, positive, and neutral) in YouTube dataset, and seven sentiments (neutral, surprise, sadness, happiness, fear, disgust, and anger) in SAVEE dataset. Further, the MALO-LSTM model performance is validated in light of accuracy, f-score, sensitivity, MCC, and specificity.

This article is prepared as follows: some existing research articles on the topic "multimodal sentiment analysis" are reviewed in section 2. Theoretical explanations and experimental analysis of the proposed MALO-LSTM model are denoted in the sections 3 and 4. Conclusion of this study is indicated in section 5.

2. Related works

M.G. Huddar, *et al*, [13] used multimodal corpus of sentiment intensity dataset and interactive motional dyadic motion capture dataset for emotion detection and multimodal sentiment analysis. Initially, Z-score standardization approach was employed on the audio modality for voice intensity threshold and voice normalization, and then 6392 feature vectors were extracted from the audio signals: arithmetic mean, pitch, standard deviation, amplitude mean, voice intensity, root quadratic mean, etc. Further, word2vec and three-dimensional convolutional neural network (CNN) models were applied for textual and visual feature extraction. The extracted feature vectors were given as the input to the softmax classifier for emotion detection, and multimodal sentiment analysis. But, the presented model considered only modality related contextual feature vectors that may degrade the performance in large datasets.

H. Najadat, and F. Abushaqra [14] used Arabic YouTube video to perform multimodal sentiment analysis. Initially, pitch, intensity, voice power and voice energy features were extracted from the audio modality. Next, the average percentage of the eye opening, and smile were calculated from the visual modality and then feature level fusion was accomplished to combine the extracted features of audio and visual modalities. Lastly, the obtained features were fed to the artificial neural network (ANN), decision tree and K nearest neighbor (KNN) for data classification. The semantic space between the extracted local features were high that leads to poor classification, while applying deep learning classification techniques.

J. Yu, *et al*, [15] developed a model: entity sensitive attention and fusion network (ESAFN) for classifying the multimodal sentiments. In this literature, the presented ESAFN model combines visual and textual modalities to attain better sentiment classification. The computation complexity of the presented model was higher, while utilizing more hand crafted feature extraction techniques.

S. Seo, *et al*, [16] presented a new heterogeneous modality transfer learning model that uses textual modality as a source to enhance the performance of visual-audio sentiments. The presented model utilizes an adversarial learning approach and a decoder for diminishing the semantic space between the target and source modalities. The simulation investigation showed that the presented model obtained a significant performance in multimodal sentimental analysis related to the conventional models. However, the class imbalance was a main issue faced by the researchers, while developing the transfer learning model.

S. Poria, *et al*, [17] implemented a multimodal affective system based on multiple kernel learning (MKL) and CNN for an effective sentiment analysis. The developed hybrid model performance was validated on YouTube dataset and the obtained results significantly outperformed the existing models in light of accuracy and computation time. As a future extension, a pre-processing technique was required to avoid the problems like multiple topic and introductory title.

Y. Zhang, *et al*, [18] presented a new quantum inspired system based on quantum interference inspired multimodal decision fusion (QIMF) approach for sentiment analysis. In addition to this, S. Angadi and V. S. Reddy, [19] initially utilized linear predictive coefficient, flux features, spectral centroid, and Mel frequency cepstral coefficients for extracting audio features from the YouTube dataset. Further, the descriptor level features and latent semantic analysis were used for extracting visual and textual features from the videos. After performing feature level fusion, the reliefF model was employed for choosing the

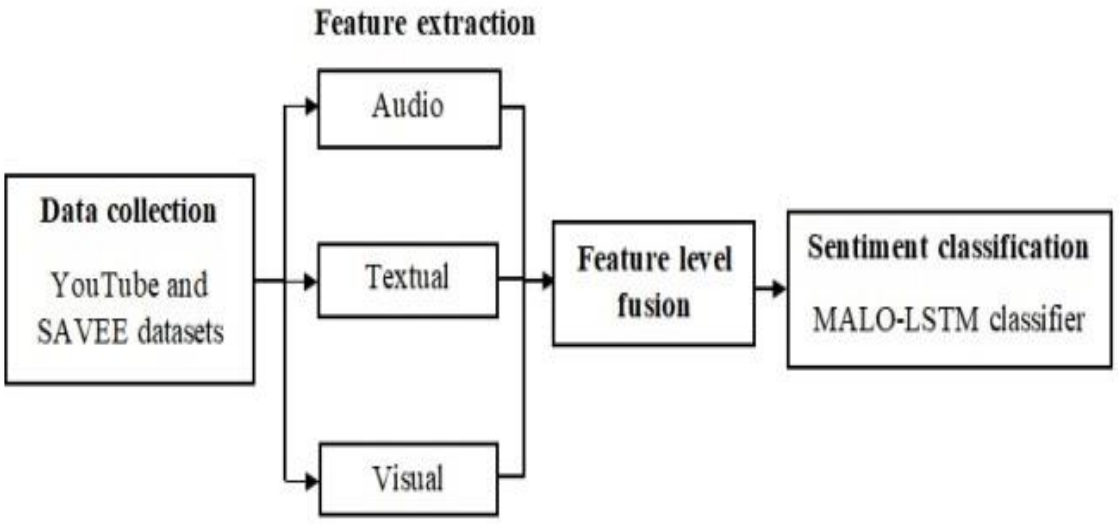

Figure. 1 Workflow of MALO-LSTM model

Table 1. Data statistics of SAVEE dataset

Number of emotions	Total files	Contribution rate (%)
Neutral	120	25
Surprise	60	12.5
Sadness	60	12.5
Happiness	60	12.5
Fear	60	12.5
Disgust	60	12.5
Anger	60	12.5

discriminative features for better classification. The discriminative features were fed to the random forest for sentimental classification such as neutral, positive and negative. However, the computational complexity of the presented model was high, so it may not be applicable in an effective real time automated system.

S. Bairavel and M. Krishnamurthy, [20] combined grass bee optimizer and multilayer perceptron-based neural network for feature selection and sentiment classification. Experimental evaluation revealed that the presented model obtained better classification accuracy with limited computational time.

Similarly, S. Kwon, *et al*, [21] has combined iterative neighborhood component analysis (INCA) and softmax classifier for feature selection and classification. However, the presented feature selection techniques such as grass bee optimizer and INCA were incapable in dealing with complex sentences. In order to overcome the above stated problems, a new model: MALO-LSTM is developed in this article for better sentiment classification.

3. Methodology

In the multimodal sentiment analysis, the proposed MALO-LSTM model includes three main steps such as

Data collection: YouTube and SAVEE dataset,
Feature extraction:
Audio: spectral centroid features, spectral flux features and short term energy,
Visual: AlexNet features, and
Text: LDA and LSA and
SentimentClassification: MALO-LSTM model.
The workflow of the proposed MALO-LSTM model is specified in Fig. 1.

3.1 Dataset description

The efficiency of the proposed MALO-LSTM model is tested on YouTube and SAVEE datasets. The YouTube dataset comprises of 47 video samples on dissimilar topics like product review and political opinions. In the YouTube dataset, 20 video samples are related to female speakers and 27 video samples are related to male speakers, and the age range of the speakers varied from 14 to 60 [22]. Each video sample has the pixel size of 360×480 with the duration of 2 to 5 minutes in MP4 format. In addition, the SAVEE dataset comprises of 480 utterances with dissimilar emotions that are recorded from 4 male actors. In the SAVEE dataset, high quality audio-visual equipment's are utilized to record the audio files in the visual media lab [23]. The data statistics of SAVEE dataset are stated in Table 1. The sample frames of YouTube and SAVEE datasets are represented in Fig. 2.

(a)

(b)

Figure. 2 Sample frames: (a) YouTube dataset and (b) SAVEE dataset

Table 2. Design configuration of AlexNet model

Number of layers		Design configuration
Fully connected	1	2092 nodes with ReLU
	2	2092 nodes with ReLU
	3	180 nodes with ReLU
Convolutional	1	150 filters in size 7×7 with MPO
	2	150 filters in size 5×5 with MPO
	3	270 filters in size 5×5 with MPO
	4	270 filters in size 3×3 with MPO
	5	180 filters in size 3×3 with MPO

3.2 Feature extraction

After converting the video sequences into frames, feature extraction is done for individual modalities (audio, visual, and textual). Firstly, spectral centroid features, spectral flux features and short term energy are employed for extracting feature vectors from the audio signals. The energy associated with the short term region of speech signal is named as short term energy. Usually, the speech signals are categorized into three types such as silence, unvoiced and voiced regions based on the nature of production. In the short term energy technique, the energy associated with voiced region is higher related to other two regions, where this technique is commonly used for unvoiced, silence and voiced speech classification [24]. In signal processing, the spectral flux features are utilized for estimating the variations in power-spectrum of the speech signal by relating its power-spectrum of one frame with other frames [25]. In addition, the spectral centroid features are utilized for measuring the signal's spectrum characteristics [19]. The spectral centroid is determined as the mean/average of signal-frequencies that is mathematically defined in Eq. (1).

$$Spectral\ centroid = \frac{\sum_{n=1}^{N} g(n)z(n)}{\sum_{n=1}^{N} z(n)} \quad (1)$$

Where, $z(n)$ indicates weighted or magnitude frequency value, $g(n)$ states center frequency with bin n, and N indicates number of samples. Further, AlexNet model is used to extract deep features from the visual modality [26]. In AlexNet model, the video frames are re-sized to the pixel value of227×227, where the re-sized frames are fed to the input layer for performing feature extraction. The AlexNet model contains 3 fully connected layers and 5 convolutional layers, where each layer is followed by neither max pooling operation (MPO) nor rectifier linear unit (ReLU) activation function. In this scenario, the deep feature vectors are extracted from the last fully connected layer along with softmax classifier. The AlexNet model's design configuration is depicted in Table 2.

Lastly, LSA and LDA techniques are used to extract textual features from the YouTube and SAVEE datasets. The LSA technique is used to represent and extract the contextual usage of texts by using statistical computations. In other words, the LSA technique evaluates and determines the pattern in un-structured text collection and also effectively analysis the relation between the texts. Correspondingly, LDA is a probabilistic topic technique, which is used for discovering the latent topics from the documents [27]. The LDA technique describes every document with a probability distribution function p, which is determined by the Eqs. (2), (3) and (4).

$$p(\aleph|\pi) = \frac{\Gamma(\sum_{i=1}^{k} \pi_i)}{\prod_{i=1}^{k} \Gamma(\pi_i)} \aleph_1^{\pi_1 - 1} \dots \dots \aleph_k^{\pi_k - 1} \quad (2)$$

$$p(\aleph, x, y|\pi, \mu) = p(\aleph|\pi) \prod_{n=1}^{N} p(x_n|\aleph)\, p(y_n|x_n, \beta) \quad (3)$$

$$p(D|\pi, \mu) = \prod_{d=1}^{M} \int p(\aleph_d|\pi) \times \left(\prod_{n=1}^{N_d} \sum_{x_{dn}} p(x_{dn}|\aleph_d) p(y_{dn}|x_{dn}, \mu)\right) d\aleph_d \quad (4)$$

Where, $\aleph$ specifies document level topic vectors, μ indicates topics, Γ denotes gamma function, N represents number of samples/text, π states dirichlet parameter, M indicates text review, x denotes topic assignment upto k^{th} text, y indicates observed text, and D states Dirichlet distribution. The extracted features from textual, audio and visual modalities are 41, 24 and 4096, respectively. The total extracted features are fed to the MALO-LSTM model for sentiment classification.

3.3 Classification

In the multimodal sentiment analysis, different modalities data need to be processed in order to attain better results. Therefore, the LSTM network is the best choice for sentiment classification by considering this aspect. The LSTM network generally composed of a series of units that stores temporal quasi-periodic features in order to extract both short and long dependencies. Generally, the LSTM network consists of input gate in_t, cell c_t, forget gate f_t, and output gate ou_t [28] that are mathematically depicted in the Eqs. (5), (6), (7) and (8).

$$in_t = \sigma(W_{inh}h_{t-1} + W_{ina}a_t + b_{in}) \quad (5)$$

$$f_t = \sigma\left(W_{fh}h_{t-1} + W_{fa}a_t + b_f\right) \quad (6)$$

$$c_t = f_t \odot c_{t-1} + in_t \odot \tanh(W_{ch}h_{t-1} + W_{ca}a_t + b_c) \quad (7)$$

$$ou_t = \sigma(W_{ouh}h_{t-1} + W_{oua}a_t + b_{ou}) \quad (8)$$

Where, $a_t = A[t,.] \in \mathbb{R}^F$ specifies temporal quasi-periodic features at time step t, W and b states work coefficients, $\sigma(.)$ specifies sigmoid activation-function, $tanh\,(.)$ denotes hyperbolic tangent activation-function. The LSTM unit's output value h_{t-1} is determined in Eq. (9).

$$h_t = ou_t \odot \tanh(c_t) \quad (9)$$

During the training and testing process, the cell state $\{c_t|t = 1,2,..T|\}$ learns the information of a_t based on the dependency relationship. Lastly, the extracted features are indicated by the last LSTM unit output h_T. The hyper-parameters selected by using MALO algorithm are; learning rate is 0.001, minimum batch size is 25, gradient threshold value is one, maximum epoch is 120, and the hidden layers are 3 (layer 1: 150 units, layer 2: 150 units, and layer 3: 100 units).

The MALO algorithm mimics the behaviour of antlions for solving the optimization problems. As similar to other population-based optimization algorithms, the MALO algorithm approximates the optimal solutions by employing a set of random solutions [29]. In the MALO algorithm, a set of random solutions is enhanced based on the interactions between ants and antlions. Totally, there are two populations in the MALO algorithm: set of antlions and set of ants. The steps involved in the MALO algorithm are given as follows:

Step 1: Initialize the ant set with random values.

Step 2: Then, estimate the fitness value of initialized ant set utilizing an objective function in every iteration.

Step 3: In the search space, the ant's moves around the antlions utilizing random walks.

Step 4: In the 1st iteration, the antlions are considered to be on the ant's location, and then re-locate the ant's position in the next iterations, if the ant's become better.

Step 5: An antlion is allocated to every ant. If the ant becomes fitter, the respective antlion updates its position.

Step 6: The antlion is selected from the solution with the least populated neighborhood.

Step 7: If the archive is full, the solutions with most-populated neighborhood are eliminated from the archive in order to obtain optimal solutions. The parameter specifications of MALO algorithm are stated as follows: number of iterations is 120, number of runs is 15, and population size is 100. After recognizing the individual's emotion, Indian art music is recommended as per the reference [30].

The experimental investigation of the MALO-LSTM model is given in the next section.

4. Experimental results

In the multimodal sentiment analysis, the proposed MALO-LSTM model is simulated utilizing MATLAB 2018a software tool on a computer with i7 processor, 16 GB random access memory, and windows 10 operating system. In this article, the effectiveness of the proposed MALO-LSTM model is validated by comparing its performance with different optimization techniques and some benchmark models like MKL-CNN [17], reliefF-random forest [19], and INCA-softmax classifier [21] on YouTube, and SAVEE datasets. The performance of MALO-LSTM model is investigated in light of accuracy, f-score, sensitivity, MCC, and specificity. The f-score is defined as a harmonic mean of recall and precision value that is mathematically denoted in Eq. (10). The sensitivity calculates the percentage of actual positive's that are correctly classified, and it is mathematically determined in Eq. (11).

$$F - score = \frac{2TP}{2TP+FP+FN} \times 100 \quad (10)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \quad (11)$$

The MCC is one of the important performance measures in machine learning, which is used to

Table 3. Performance valuation of MALO-LSTM model with different modalities on YouTube dataset

MALO-LSTM model					
Modalities	**Accur-acy (%)**	**Sensiti-vity (%)**	**Specific-ity (%)**	**F-score (%)**	**MCC (%)**
Textual	94.06	95.24	97.62	92.32	91.65
Audio	86.90	95.43	82.14	86.75	81.01
Visual	94.05	94.43	96.43	93.81	92.41
Visual and textual	92.86	93.43	91.43	91.83	90.16
Audio and textual	94.05	94.64	92.86	91.26	86.81
Visual and audio	95.24	94.64	93.10	94.94	92.50
Textual, audio and visual	**98.62**	**96.43**	**99.08**	**96.55**	**95.87**

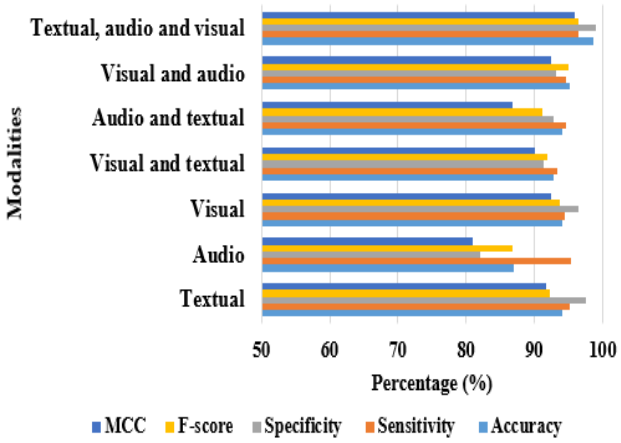

Figure. 3 Graphical presentation of MALO-LSTM model with different modalities on YouTube dataset

Table 4. Performance valuation of LSTM network with different optimizers on YouTube dataset

LSTM network					
Optimi-zers	**Accur-acy (%)**	**Sensi tiv-ity (%)**	**Specifi c-ity (%)**	**F-score (%)**	**MCC (%)**
GWO	89.29	88.01	90.86	91.30	92.69
GOA	94.05	91.07	93.58	92.03	88.26
WOA	97.62	95.21	96.43	93.43	93.64
MALO	**98.62**	**96.43**	**99.08**	**96.55**	**95.87**

measure the quality of MALO-LSTM model and the accuracy performance measure calculates the percentage of correctly classified classes. The mathematical expressions of MCC and accuracy are defined in the Eqs. (12) and (13). The specificity estimates the percentage of actual negative's that are correctly classified, and it is mathematically denoted in Eq. (14). Where, TP, TN, FP and FN are indicated as true positive, true negative, false positive and false negative.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}} \times 100 \quad (12)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (13)$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \quad (14)$$

4.1 Quantitative study on YouTube dataset

Here, the effectiveness of the MALO-LSTM model is validated on the YouTube dataset in terms of f-score, sensitivity, MCC, accuracy and specificity. Among 210 video samples, 168 samples are used for MALO-LSTM training, and 42 samples are used for MALO-LSTM testing with five-fold cross validation. By viewing Table 3, the proposed MALO-LSTM model effectiveness is validated by tilizing different modalities. As mentioned in Table 3, the MALO-LSTM model attained better results in sentiment analysis by combining all three modalities like textual, audio, and visual. On the YouTube dataset, the MALO-LSTM model obtained a maximum f-score of 96.55%, sensitivity of 96.43%, MCC of 95.87%, accuracy of 98.62%, and specificity of 99.08%, where the achieved results are better related to individual modalities and combined modalities. Graphical presentation of MALO-LSTM model with different modalities on YouTube dataset is specified in Fig. 3.

In Table 4, the effectiveness of LSTM network is validated with dissimilar optimizers such as grey wolf optimization (GWO) algorithm, grasshopper optimization algorithm (GOA), whale optimization algorithm (WOA), and MALO algorithm by means of f-score, sensitivity, accuracy, MCC, and specificity. By inspecting Table 4, the LSTM network with MALO algorithm achieved significant performance in sentiment analysis related to comparative optimization algorithms. The optimum hyper-parameter selection in LSTM network is important, because it directly controls the behaviour of testing and training. Additionally, the selection of optimum hyper-parameters easily manages a large set of experiments for hyper-parameter tuning. The

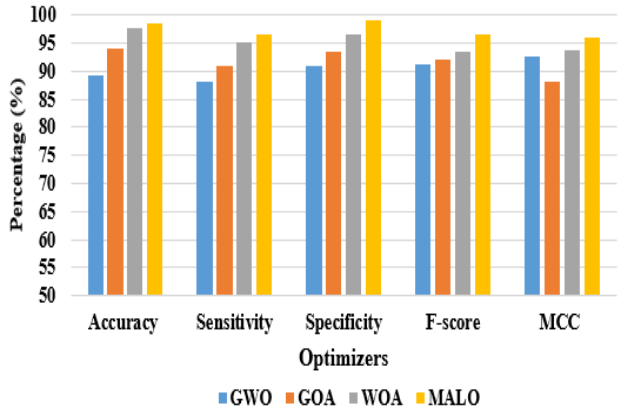

Figure. 4 Graphical presentation of LSTM network with different optimizers on YouTube dataset

Table 5. Performance evaluation of MALO-LSTM model with different modalities on SAVEE dataset

MALO-LSTM model					
Modalities	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F-score (%)**	**MCC (%)**
Textual	92.81	98.61	97.10	96.15	95.64
Audio	83.93	89.61	91.67	91.48	90.20
Visual	94.64	96.53	95.10	90.60	89.42
Visual and textual	89.29	88.61	83.33	86.96	85
Audio and textual	89.88	95.14	91.67	83.05	80.42
Visual and audio	92.26	95.31	87.50	90.86	90.13
Textual, audio and visual	**98.81**	**99.31**	**100**	**98**	**97.70**

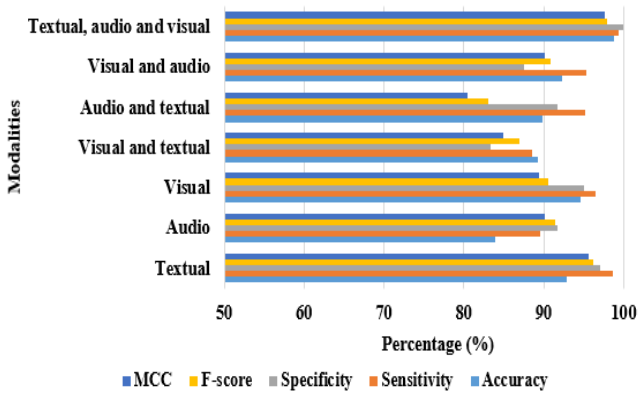

Figure. 5 Graphical presentation of MALO-LSTM model with different modalities on SAVEE dataset

graphical illustration of LSTM network with different optimizers on YouTube dataset is specified in Fig. 4.

4.2 Quantitative study on SAVEE dataset

Correspondingly on SAVEE dataset, the proposed MALO-LSTM model obtained a better performance in sentiment analysis by combining all three modalities. The proposed MALO-LSTM model achieved f-score of 98%, sensitivity of 99.31%, MCC of 97.70%, accuracy of 98.81%, and specificity of 100% in multimodal sentiment analysis, where the obtained results are better related to other combined modalities, as indicated in Table 5. The graphical presentation of MALO-LSTM model with different modalities on SAVEE dataset is specified in Fig. 5. In this study, the feature level fusion is carried out to combine all the feature vectors extracted from each modality (visual, textual, and audio), which is eventually given as the input to a classification methodology. One of the major problems in feature level fusion is the assimilation of the heterogeneous feature vectors, which is significantly overcome by employing the proposed MALO-LSTM model.

As same as Table 4, the performance of LSTM network is validated with different optimizers in light of f-score, sensitivity, accuracy, MCC, and specificity on SAVEE dataset, where the obtained results are depicted in Table 6. Compared to other existing optimizers like GWO, GOA and WOA, the combination: LSTM network with MALO algorithm attained better result in sentiment analysis on SAVEE dataset. The LSTM network uses a large range of parameter values such as input biases, learning rate and output biases, so there is no need of any fine adjustments for classification. This process reduces the complexity of proposed MALO-LSTM model to linear $O(N)$, where N states input data size, and O specifies order of magnitude. The graphical presentation of LSTM network with different optimizers on SAVEE dataset is indicated in Fig. 6.

Table 6. Performance evaluation of LSTM network with different optimizers on SAVEE dataset

LSTM network					
Optimizers	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F-score (%)**	**MCC (%)**
GWO	90.48	95.31	87.50	91.30	90.04
GOA	89.29	98	83.33	90.68	89.97
WOA	92.26	97.31	95.83	95.83	95.14
MALO	**98.81**	**99.31**	**100**	**98**	**97.70**

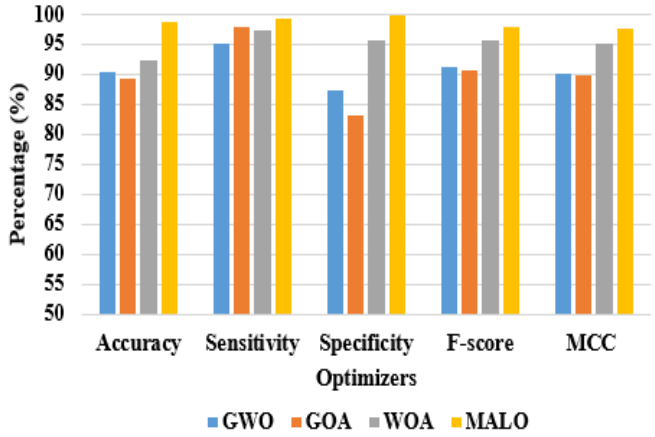

Figure.6 Graphical presentation of LSTM network with different optimizers on SAVEE dataset

Table 7. Comparative study between the existing and proposed MALO-LSTM model

Models	Dataset	Accuracy (%)	Sensitivity (%)
MKL-CNN [17]	YouTube	88.60	87.36
ReliefF-random forest [19]	YouTube	94.01	91
INCA-softmax classifier [21]	SAVEE	95	81
MALO-LSTM model	**YouTube**	**98.62**	**96.43**
	SAVEE	**98.81**	**99.31**

4.3 Comparative study

The comparative valuation between the previous and proposed MALO-LSTM model is stated in Table 7. S. Poria, *et al*, [17] combined MKL and CNN model for an effective sentiment detection. In this work, the MKL-CNN model performance is tested on YouTube dataset, where MKL-CNN model achieved 88.60% of classification accuracy, and 87.36% of sensitivity on YouTube dataset. S. Angadi and V.S. Reddy, [19] used reliefF algorithm and random forest classifier for multimodal sentiment analysis. After extracting the multiple feature vectors, the reliefF algorithm was utilized to choose the discriminative feature values, and then fed to the random forest for classification. As represented in the resulting section, the presented reliefF-random forest model attained 94.01% of classification accuracy and 91% of sensitivity on YouTube dataset. S. Kwon, *et al*, [21] combined INCA and softmax classification technique for an effective multimodal sentiment analysis. Hence, the presented INCA-softmax model obtained 95% of classification accuracy and 81% of sensitivity on SAVEE dataset. Compared to prior research studies, the MALO-LSTM model obtained good performance in multimodal sentiment analysis on YouTube dataset and SAVEE dataset in light of sensitivity and accuracy. In this article, the MALO-LSTM model utilizes active feature vectors for better classification with low computational complexity. The obtained experimental results reveals that the MALO-LSTM model effectively address the issues mentioned in the literatures [13 - 16, 19, 21].

5. Conclusions

In this article, a new MALO-LSTM model is proposed for multimodal sentiment analysis and then recommends the Indian art music raga (Carnatic music) based on the individual's sentiments. This procedure reduces depression, sleep better, increases the learning strength, improves running performance, reduces stress, and improves health. The MALO-LSTM model majorly includes two steps: feature extraction and sentiment classification. For feature extraction, several statistical and deep learning techniques are employed to extract feature vectors from the textual, visual and audio modalities on YouTube and SAVEE datasets. The obtained discriminative feature vectors are fed to the MALO-LSTM model for sentiment classification, where seven sentiments (neutral, surprise, sadness, happiness, fear, disgust, and anger) in SAVEE, and three sentiments (negative, positive, and neutral) in YouTube dataset. In this study, the efficiency of MALO-LSTM model is validated in light of accuracy, f-score, sensitivity, MCC, and specificity. As seen in the comparative analysis section, the MALO-LSTM model achieved 98.62% and 98.81% of classification accuracy on both YouTube and SAVEE datasets, where the attained results are better related to comparative models. However, the present feature level feature technique is simpler, so as a future extension, it can be substituted by training a classifier.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, methodology, software, formal analysis, resources, data curation, and writing-original draft preparation, writing-review, editing: Sri Raman Kothuri, and Supervision: N R RajaLakshmi.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. Further, the MATLAB code of the new model/method used in this analysis will be made available through matlab central once this research article is published.

https://in.mathworks.com/matlabcentral/profile/authors/11633819.

References

[1] A. Agarwal, A. Yadav, and D. K. Vishwakarma, "Multimodal sentiment analysis via RNN variants", In: *Proc of 2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, pp. 19-23, 2019.

[2] Y. Li, K. Zhang, J. Wang, and X. Gao, "A cognitive brain model for multimodal sentiment analysis based on attention neural networks", *Neurocomputing*, Vol. 430, pp. 159-173, 2021.

[3] H. N. Tran and E. Cambria, "Ensemble application of ELM and GPU for real-time multimodal sentiment analysis", *Memetic Computing*, Vol.10, No. 1, pp. 3-13, 2018.

[4] P. D. Mahendhiran and S. Kannimuthu, "Deep learning techniques for polarity classification in multimodal sentiment analysis", *International Journal of Information Technology & Decision Making*, Vol. 17, No. 03, pp. 883-910, 2018.

[5] T. Zhou, J. Cao, X. Zhu, B. Liu, and S. Li, "Visual-Textual Sentiment Analysis Enhanced by Hierarchical Cross-Modality Interaction", *IEEE Systems Journal*, 2020.

[6] Z. Zhao, H. Zhu, Z. Xue, Z. Liu, J. Tian, M. C. H. Chua, and M. Liu, "An image-text consistency driven multimodal sentiment analysis approach for social media", *Information Processing & Management*, Vol.56, No.6, p. 102097.

[7] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling", *Knowledge-Based Systems*, Vol. 161, pp. 124-133, 2018.

[8] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image-text sentiment analysis via deep multimodal attentive fusion", *Knowledge-Based Systems*, Vol. 167, pp. 26-37, 2019.

[9] S. Poria, E. Cambria, N. Howard, G. B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content", *Neurocomputing*, Vol. 174, pp. 50-59, 2016.

[10] A. Kumar, K. Srinivasan, W. H. Cheng, and A. Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data", *Information Processing & Management*, Vol. 57, No. 1, p. 102141, 2020.

[11] H. Wen, S. You, and Y. Fu, "Cross-modal context-gated convolution for multi-modal sentiment analysis", *Pattern Recognition Letters*, Vol. 146, pp. 252-259, 2021.

[12] W. Liao, B. Zeng, J. Liu, P. Wei, X. Cheng, and W. Zhang, "Multi-level graph neural network for text sentiment analysis", *Computers & Electrical Engineering*, Vol. 92, p. 107096, 2021.

[13] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification", *International Journal of Multimedia Information Retrieval*, Vol. 9, No. 2, pp. 103-112, 2020.

[14] H. Najadat and F. Abushaqra, "Multimodal sentiment analysis of Arabic videos", *Journal of Image and Graphics*, Vol. 6, No.1, pp. 39-43, 2018.

[15] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 429-439, 2019.

[16] S. Seo, S. Na, and J. Kim, "HMTL: Heterogeneous Modality Transfer Learning for Audio-Visual Sentiment Analysis", *IEEE Access*, Vol. 8, pp. 140426-140437, 2020.

[17] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis", *Neurocomputing*, Vol. 261, pp. 217-230, 2017.

[18] Y. Zhang, D. Song, P. Zhang, P. Wang, J. Li, X. Li, and B. Wang, "A quantum-inspired multimodal sentiment analysis framework", *Theoretical Computer Science*, Vol. 752, pp. 21-40, 2018.

[19] S. Angadi, and V. S. Reddy, "Multimodal sentiment analysis using reliefF feature selection and random forest classifier", *International Journal of Computers and Applications*, pp. 1-9, 2019.

[20] S. Bairavel and M. Krishnamurthy, "Novel OGBEE-based feature selection and feature-level fusion with MLP neural network for social media multimodal sentiment analysis", *Soft Computing*, Vol. 24, pp. 18431-18445, 2020.

[21] S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network",

International Journal of Intelligent Systems, 2021.

[22] L. P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web", In: *Proc. of 2011 13th International Conference On Multimodal Interfaces*, pp. 169-176, 2011.

[23] F. Rahdari, E. Rashedi, and M. Eftekhari, "A multimodal emotion recognition system using facial landmark analysis", *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, Vol. 43, No.1, pp. 171-189, 2019.

[24] D. Dimitriadis, A. Potamianos, and P. Maragos, "A comparison of the squared energy and Teager-Kaiser operators for short-term energy estimation in additive noise", *IEEE Transactions on signal processing*, Vol. 57, No.7, pp. 2569-2581, 2009.

[25] A. R. Choudhury, A. Ghosh, R. Pandey, and S. Barman, "Emotion recognition from speech signals using excitation source and spectral features", In: *Proc of 2018 IEEE Applied Signal Processing Conference (ASPCON)*, pp. 257-261, 2018.

[26] L. Ding, H. Li, C. Hu, W. Zhang, and S. Wang, "Alexnet feature extraction and multi-kernel learning for object-oriented classification", *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, Vol. 42, pp. 277-281, 2018.

[27] J. Hoblos, "Experimenting with latent semantic analysis and latent dirichlet allocation on automated essay grading", In: *Proc of 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE*, pp. 1-7, 2020.

[28] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework", *IEEE Access*, Vol. 6, pp. 49325-49338, 2018.

[29] S. Mirjalili, P. Jangir, and S. Saremi, "Multi-objective ant lion optimizer: a multi-objective optimization algorithm for solving engineering problems", *Applied Intelligence*, Vol. 46, No.1, pp. 79-95, 2017.

[30] S. Gulati, J. Serra, V. Ishwar, S. Sentürk, and X. Serra, "Phrase-based rāga recognition using vector space modelling", In: *Proc of 2016IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66-70, 2016.