Received: 25 June 2019 | Revised: 2 December 2019 | Accepted: 29 December 2019

DOI: 10.1111/coin.12274

ORIGINAL ARTICLE

Computational Intelligence WILEY

Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification

Mahesh G. Huddar[1] | **Sanjeev S. Sannakki**[2] | **Vijay S. Rajpurohit**[2]

[1]Computer Science and Engineering, Hirasugar Institute of Technology, Belagavi, Karnataka, India

[2]Computer Science and Engineering, Gogte Institute of Technology, Belagavi, Karnataka, India

Correspondence
Mahesh G. Huddar, Computer Science and Engineering, Hirasugar Institute of Technology, Nidasoshi, Belagavi, Karnataka, India.
Email: mailtomgh1@gmail.com

Abstract

The availability of the humongous amount of multimodal content on the internet, the multimodal sentiment classification, and emotion detection has become the most researched topic. The feature selection, context extraction, and multi-modal fusion are the most important challenges in multimodal sentiment classification and affective computing. To address these challenges this paper presents multilevel feature optimization and multimodal contextual fusion technique. The evolutionary computing based feature selection models extract a subset of features from multiple modalities. The contextual information between the neighboring utterances is extracted using bidirectional long-short-term-memory at multiple levels. Initially, bimodal fusion is performed by fusing a combination of two unimodal modalities at a time and finally, trimodal fusion is performed by fusing all three modalities. The result of the proposed method is demonstrated using two publically available datasets such as CMU-MOSI for sentiment classification and IEMOCAP for affective computing. Incorporating a subset of features and contextual information, the proposed model obtains better classification accuracy than the two

standard baselines by over 3% and 6% in sentiment and emotion classification, respectively.

KEYWORDS

bidirectional LSTM, contextual information, evolutionary computing, feature selection, multimodal fusion

1 | INTRODUCTION

The increased usage of smartphones and the availability of affordable internet facilities led the user to post an audio-visual review on social media networks.[1] Hence, the multimodal sentiment classification and affective computing have become more popular among the researchers. The industries are using this large amount of information to increase their revenue. They perform sentiment analysis and emotion detection to understand the mood of customers towards their service, product, complaints, suggestions, etc. The advantage of analyzing multimodal content over the textual data is the availability of audio-visual content, which greatly improves the overall performance of the system. In multimodal sentiment classification and affective computing, features are extracted from individual modalities such as text, acoustic, and visual.[2] The main issues in multimodal sentiment analysis and affective computing are feature selection and multimodal fusion. The features extracted from individual modalities are of high dimensional in nature and many of these features are redundant and irrelevant. Due to this, the training and testing process takes more time and many times degrades the training accuracy. This issue can be dealt with logically by selecting important features from individual modalities. The evolutionary computing based models such as genetic algorithm (GA), particle swarm optimization (PSO) and greedy search based genetic algorithm (GGA) are built for selecting a subset of features.

The multimodal data is divided into small segments called utterances and each video is a sequence of utterances. The effective polarity or emotion of an utterance may depend on the outcome of neighboring utterances (ie, contextual knowledge). For multimodal fusion, either early (feature level) fusion[3] or late (decision level) fusion[4] techniques are used. These simplistic models will not extract the contextual information among the utterances. Also, the audio-visual data may contain redundant and irrelevant information, which cannot be handled by these traditional fusion techniques. These issues are addressed by selecting a subset of features and using a multimodal contextual fusion technique for fusing the information from multiple modalities.

The contributions of the proposed model are,

- Designing evolutionary computing based models such as GA, PSO, and GGA for feature selection.
- Early fusion, late fusion, and recurrent neural network-based models restrict the extraction of contextual information among the neighboring utterances. Hence, bidirectional long-short-term-memory (biLSTM) based model is designed to extract the context among utterances at multiple levels and multimodal fusion for fusing information from multiple modalities.

- The performance of the proposed model was demonstrated using two publically available datasets. The results show that the proposed model achieves higher accuracy than the standard baselines. Also, the time required for training the model with a subset of selected features is relatively less compared to all features.

The structure of the remaining article follows: the important recent work in feature selection methods, multimodal sentiment analysis, affective computing, and traditional multimodal fusion techniques is discussed in Section 2. The proposed multi-level feature selection, weighted feature ensemble, and multimodal fusion techniques are discussed in detail in Section 3. Section 4 gives the experimental results and analysis of the proposed model. Finally, the future work in multimodal sentiment classification and emotion detection is presented and concludes the article in Section 5.

2 | LITERATURE REVIEW

The multimodal data from the internet is analyzed for understanding the user's sentiment or emotion towards a product or a service.[5] The sentiment analysis and affective computing have a lot of challenges such as subjectivity analysis,[6] aspect detection and consideration,[7] topic detection and tracking,[8] document summarization.[9] The earlier research concentrates on textual data for sentiment analysis[10] but more recently multimodal data such as textual-acoustic-visual modality were considered for sentiment classification and affective computing.[11] Usually, handcrafted features[12] or lexicons[13] or networks[14] or ontologies[15] are used in sentiment classification.

Feature selection is one of the main issues in machine learning. The aim of feature selection is to select the important features for a given problem. By removing the redundant, irrelevant and ambiguous features, the feature selection reduces the dimensionality and improves the training process of the machine learning task and also improves the classification accuracy.[16] Most of the existing feature selection models use either exhaustive search[17] or heuristic search[18] or backtracking algorithm.[19] The experimental results show that the performance of the heuristic algorithm is the same as the backtracking algorithm but heuristic search-based algorithm takes much less time. More recently evolutionary computing-based algorithms are being used to address the feature selection issue. Traditional search based methods need domain knowledge, but evolutionary computing-based algorithms work without domain knowledge. These algorithms make an assumption about the search space.

Another issue in multimodal affective computing is multimodal fusion.[20] Multimodal sentiment analysis is performed on movie reviews collected from YouTube.[21] Unigram, bigram and trigram bag of word features for textual modality, head pose, gaze direction and smile intensity of facial modality and acoustic features like low-level descriptors (LLD) were extracted. Linguistic features are trained using support vector machine and audio-visual features are trained using bidirectional long short-term memory and cross-domain Metacritic corpus was used to evaluate the proposed model.

The multimodal opinion Utterances dataset (MOUD) dataset containing video reviews in the Spanish language proposed in Reference 3 Bag-of-word representations for textual data, smile duration, facial expressions for visual modality and MFCC features for acoustic modality are extracted.[22] Early fusion was used to merge the unimodal features from individual modalities and show that the multimodal model outperforms the unimodal model. Multimodalities such as text-video-audio were used to detect sentiment intensity in Reference 23. Feature level and

TABLE 1 Comparative study on multimodal sentiment analysis and emotion detection

Ref. no.	Year	Dataset	Algorithm	Multimodal fusion	Features	Findings
Wöllmer et al[21]	2013	ICT-MMMO	SVM for linguistic analysis & BLSTM for audio-visual analysis	Feature level fusion	Bag-of-words, smile intensity, head pose, gaze direction, low-level descriptors	Performance of audio-visual combination of modality is better even if the textual modality is not considered
Pérez-Rosas et al[3]	2013	MOUD dataset (Spanish)	SVM for sentiment classification	Feature level fusion	Acoustic, linguistic, and visual features	The error rate was reduced by 10.5% compared to unimodal models
Rosas et al[22]	2013	MOUD dataset (Spanish)	SVM for classification	Feature level fusion	Bag-of-words, smile duration, head pose, gaze direction, MFCC, and low-level descriptors	Experiments showed that the multimodal model performs better than the unimodal model
Zadeh[23]	2015	CMU-MOSI	SVM for subjectivity analysis	Decision level fusion	Ngrams from the text, MFCC and peak slope from audio & FAU, FL & head pose from visual modality were extracted	Effectiveness of multimodal analysis in subjectivity recognition
Poria et al[4]	2016	YouTube dataset	SVM, ELM & ANN are used for classification	Feature/decision fusion	Visual & acoustic features using FSDK 1.7, open EAR and Concept-gram and Sentic Net-based textual features are used	The accuracy of feature-level fusion and ELM classifier better compared to decision level fusion, SVM and ANN classifier
Poria et al[24]	2017	MOUD, YouTube & ICT-MMMO	Convolutional multiple kernel earning (MKL) method	Decision fusion	Acoustic, linguistic, and visual features	The proposed method outperforms the standard baselines
Majumder et al[25]	2018	IEMOCAP, CMU-MOSI	Recurrent neural networks (RRN)	Hierarchical Fusion	CNN for textual, Open SMILE for acoustic and 3D CNN for visual features was used	Outperforms the standard baseline by over 1% in terms of classification accuracy
Pham et al[26]	2018	CMU-MOSI, ICT-MMMO, YouTube	Sequence to sequence models	Hierarchical Fusion	Acoustic, Linguistic, and Visual features	Joint representations outperform unimodal models
Proposed work		CMU-MOSI IEMOCAP	GA, PSO, GGA for feature selection, RRN for fusion	Hierarchical contextual fusion	CNN for textual, Open SMILE for acoustic and 3D CNN for visual features was used	Outperforms the standard baselines in classification accuracy with less execution time

decision level fusion techniques and support vector machines (SVM), extreme learning method (ELM) and artificial neural networks were used as a classifier in Reference 4

The textual, visual and acoustic features are extracted using the convolutional neural network (CNN) and sentiment and emotion recognition are performed using multiple-kernel learning.[24] They constructed the CMU-MOSI dataset of 93 videos collected from YouTube. Ngrams from a textual modality, MFCC and peak slope from acoustic modality and facial action units (FAU), facial landmarks (FL) and head pose from visual modality were extracted. Instead of a traditional feature level or decision level fusion, hierarchical fusion was used in Reference 25. The proposed novel fusion technique outperforms the standard baselines by over 1%. Sequence-to-sequence based model is used to understand the joint representation between multiple modalities.[26] Recently shallow fusion models such as Tensor fusion[27] model are built for multimodal sentiment classification. Table 1 shows the comparative study on multimodal sentiment and emotion classification.

Research gaps:

The existing methods use CNN for extracting textual features, Open SMILE or Open EAR for extracting acoustic features and FSDK 1.7 or 3D CNN for extracting visual features. The dimensionality of the unimodal feature vector extracted using these methods is large and may contain irrelevant and redundant features. The traditional fusion techniques use all raw features for training sentiment and emotion classification model. This makes the training machine learning task slow and may produce contradictory results. Unlike the existing methods, the proposed model extracts the subset of features from individual modalities using evolutionary computing based feature selection methods. The bidirectional LSTM is used to extract the contextual data from a subset of features. Finally, multilevel contextual fusion is used for multimodal sentiment analysis and affective computing. Finally, the model is demonstrated using two publically available standard datasets.

3 | PROPOSED METHODOLOGY

This section discusses the proposed feature selection models and multimodal contextual fusion in detail. The Overview of the proposed model is:

- First, textual, acoustic and visual features are extracted at utterance level.
- The evolutionary computing based feature selection model is used for selecting a subset of features of individual modalities.
- Unimodal contextual features are extracted using bi-LSTM
- Two-two modalities are fused at a time to get the bimodal features.
- A bimodal subset of features is selected and bimodal contextual features are extracted using bi-LSTM.
- All modalities are fused to get trimodal feature vector
- A trimodal subset of features is selected and contextual features are extracted using bi-LSTM.
- The utterance-level multimodal contextual feature vector with a subset of features is finally used for sentiment and emotion classification.

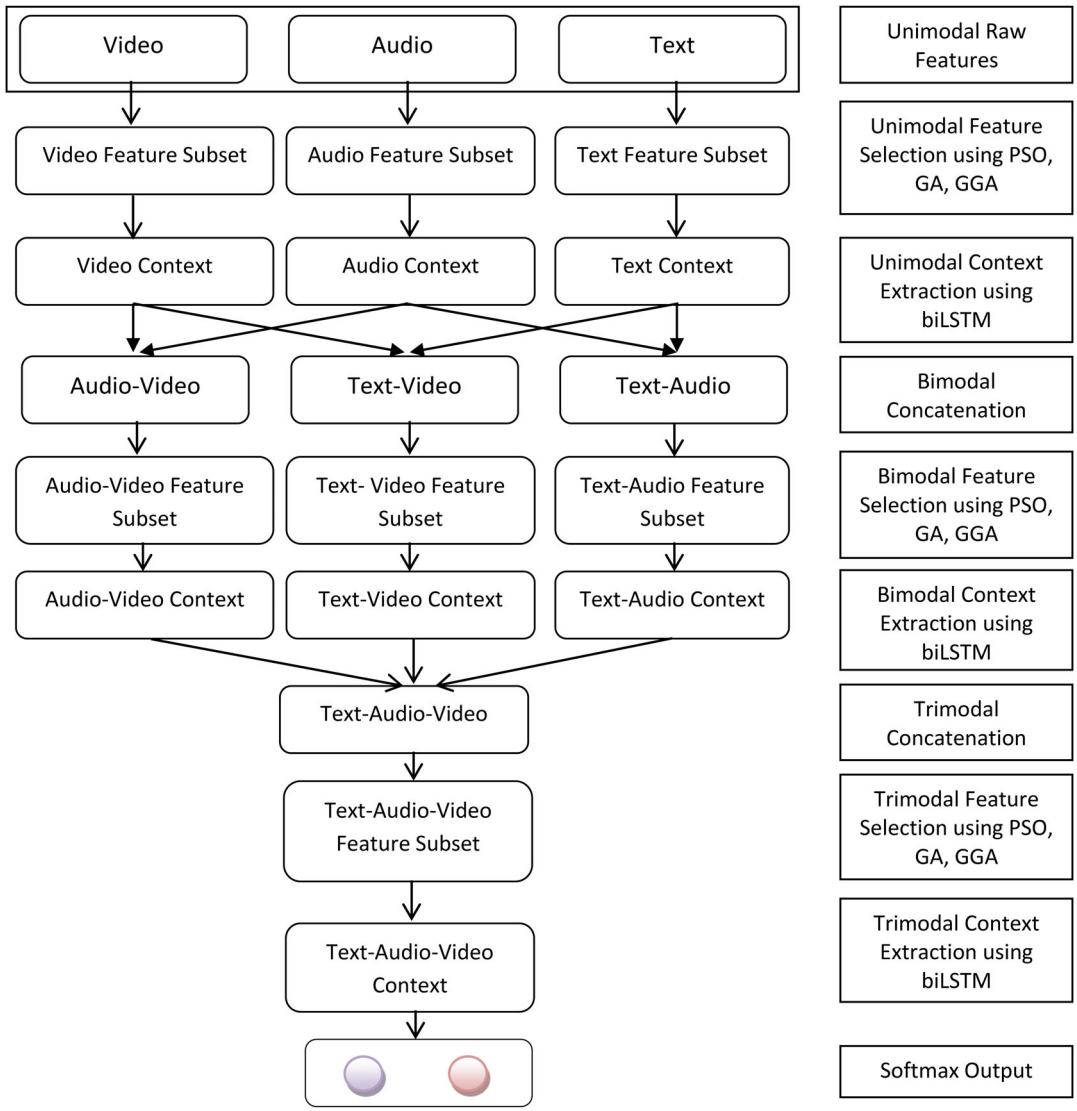

FIGURE 1 Proposed multi-level feature optimization and multimodal contextual fusion methodology [Color figure can be viewed at wileyonlinelibrary.com]

The proposed multi-level feature optimization and multimodal contextual fusion methodology is shown in Figure 1 and discussed in detail in further sections.

The subset of features is selected using evolutionary computing based feature selection methods. Let F be the feature set, A be the accuracy and T be the execution time with all features, then the objective function of feature optimization is to select the optimal subset of features F_s,

$$F_s = \text{featureSelect}(F, A, T), \tag{1}$$

such that $A_s > A$ and $t_s < t$, where A_s is the accuracy and t_s is the execution time with a subset of features.

3.1 | Dataset used

The proposed models are evaluated on two publically available multimodal datasets such as IEMOCAP[28] and CMU-MOSI[29] for multimodal sentiment analysis and emotion classification, respectively.

3.1.1 | IEMOCAP

The IEMOCAP dataset contains two way recorded dyadic dialogs between two speakers. The dataset is a collection of 302 videos and each video is divided into multiple small segments or utterances. Each segment or utterance is annotated by multiple assessors for the presence of nine different emotions such as angry, excitement, fear, neutral, surprised, frustrated, happiness, sadness, and disappointment. To be in line with the recent research, the study considers four emotion states such as angry, happiness, neutral and sadness for the experimentation. To evaluate the proposed model the dataset is divided into train-test sets (such that the splits are speaker-independent). Table 2 shows the train-test split of the IEMOCAP dataset.

3.1.2 | CMU-MOSI

The CMU-MOSI dataset is a collection of 93 opinion videos collected from YouTube. The dataset is divided into 2199 small opinion segments with an average of 12 words per segment and 4.2 seconds of segment length. Each of the segments is rigorously annotated by 5 assessors with scores in the range of $[-3, +3]$, with -3 and $+3$ being the extremely negative and extremely positive sentiment, respectively. The voted average of five assessors is taken as the sentiment for each segment. Similar to the IEMOCAP dataset the experiments are conducted on positive and negative sentiment utterances. To assess the model the dataset is divided into train-test sets and Table 3 shows the train-test split of the CMU-MOSI dataset.

3.2 | Feature extraction

The feature extraction from multiple modalities and features optimization are the important requirements of the proposed method. This section describes the utterance or segment level feature extraction from textual, audio and video modalities.

TABLE 2 Train-test split of IEMOCAP dataset

	Happy	Angry	Sad	Neutral
Train	1194	933	839	1324
Test	433	157	238	380

TABLE 3 Train-test distribution of CMU-MOSI dataset

	Positive	Negative
Train	709	738
Test	467	285

3.2.1 | Textual feature extraction

Each video utterance is transcribed to get the textual transcription. Convolutional Neural Network (CNN) model is used to extract the utterance level textual features.[30] The word2vec vectors representation is used to extract the context from the textual data. The number of words in each utterance is different; hence, either an utterance is truncated or padded with null data to prepare the uniform feature vector. The three convolutional layers are used to process the wor2vec vector representation of utterances. Filters of size 4, 3, 2 and 50, 75 and 100 feature maps are used in the three convolutional layers, respectively. Window size 2×2 with the max-pooling operation is used in all three layers. A fully connected layer with 600 computational units with ReLU[31] activation function is used between the convolutional layers and the softmax classifier. The output of the CNN is considered as the textual features.

Let t_i is the textual feature vector of i^{th} utterance and n is the number of utterances then, the textual feature vector t of the dataset is represented as,

$$t =< t_1, t_2, t_3, \ldots \ldots, t_n > . \tag{2}$$

3.2.2 | Audio feature extraction

The OpenSMILE[32] is a cross-platform audio analysis toolkit used for speech preprocessing and audio feature extraction in real-time. The utterance level audio features such as Mel-frequency cepstral coefficients (MFCCs), voice intensity, pitch, root quadratic mean, skew-ness, amplitude mean, arithmetic mean, SD, quartiles ranges, interquartile ranges, and linear regression slope are extracted at a frame rate of 30 Hz and a sliding window of size 100 ms.

Let a_i is the audio feature vector of i^{th} utterance and n is the number of utterances then, the audio feature vector a of the dataset is represented as,

$$a =< a_1, a_2, a_3, \ldots \ldots, a_n > . \tag{3}$$

3.2.3 | Visual feature extraction

In the recent literature object detection, human action recognition and classification task are successfully addressed using the 3D-CNN models.[33,34] The results show that the 3D-CNN based models outperform the standard baselines in object detection, object tracking, and video classification. The 3D-CNN model is used to extract frame-level visual features as well as temporal features across frames.[35] The study replicates the same process for extracting the temporal features across frames and visual features at the frame level from visual modality.

Let v_i is the visual feature vector of i^{th} utterance and n is the number of utterances then, the video feature vector v of the dataset is represented as,

$$v =< v_1, v_2, v_3, \ldots \ldots, v_n > . \tag{4}$$

3.3 | Feature selection

Feature selection in machine learning is the process of identifying and selecting a subset of important features, which enables the faster training of machine learning models, reduces

the over-fitting problem with higher accuracy. GA and PSO are evolutionary computing based optimization algorithms. These algorithms are widely used in many fields to address the problem of optimization. This section discusses the evolutionary computing based three feature selection methods such as GA, PSO, and G GA. In feature selection algorithm a subset of features is selected and rest are discarded in every evolution. The model is evaluated using the selected features until the stopping criterion or number of generations is reached. The binary strings are used to represent selected and unselected features. The selected features are represented with 1 and unselected features with 0.

3.3.1 | Particle swarm optimization based feature selection

The PSO based feature selection algorithm considers every possible subset of features in the population as a particle. The position of i^{th} particle (feature subset) at g^{th} generation is denoted by $X_i(t)$. For each generation, based on the fitness value or best position, the global best (*gbest*) for whole swarm and local best (*pbest*) for particle are set. The movement of particles (feature subsets) is controlled by global and local best particles. The velocity and position of the particles are updated as:

$$V_{i+1}(g+1) = w * v_i(g) + c_1 * r_1 * (lbest - v_i(g)) + c_2 * r_2 * (gbset - X_i(g)), \quad (5)$$

$$X_{i+1}(g+1) = X_i(g) + V_{i+1}(g+1), \quad (6)$$

$i = 1, 2, 3, \ldots\ldots, N$ (Number of particles)
g is the number of generations
w is the inertia weight
$0 <= c_1$ and $c_2 <= 2$ are known as cognitive and social constants.
r_1 and r_2 are uniformly distributed random numbers.
$V_i(g)$ and $X_i(g)$ are the velocity and the position of i^{th} particle at g^{th} generation.

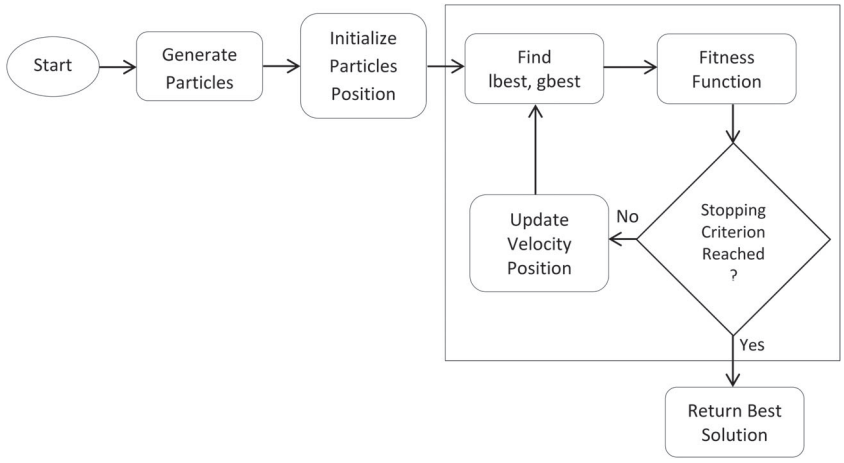

FIGURE 2 Steps in particle swarm optimization based feature selection

TABLE 4 Particle swarm optimization based feature selection

```
Input: The training data set with N attributes and K examples
Output: Feature subset
Initialize Parameters
Initialize Population
while (The stopping criterion is not met or the Number of
  generations) then
do
    for (i = 1 to number of particles)
    do
        if the fitness of X_i(g) is greater than the fitness of lbest then
        do
            update lbest = X_i(g)
        end if
        if the fitness of X_i(g) is greater than the fitness of gbest then
        do
        update gbest = X_i(g)
        end if
        Update velocity
        Update particle position
        Next particle
    end for
    Next-generation
end while
```

Through experience, the values of w and constants (c_1 and c_2) are set to 0.01 and 0.02, respectively. Figure 2 shows the basic steps in the proposed PSO algorithm for feature selection and is summarized in Table 4.

3.3.2 | Genetic algorithm based feature selection

The genetic algorithm starts with initializing the population. Crossover is performed over two randomly selected individuals (features) from the population followed by the mutation depending on the mutation rate. The fitness function gives the relative importance of each feature. A group of best features is selected based on the fitness function to form the new population for the next generation. Fixed number generations are used as the stopping criteria for the GA. Figure 3 shows the flowchart for the GA based feature selection and is summarized in Table 5.

3.3.3 | Greedy search based genetic algorithm for feature selection

In every generation, a GA based feature selection algorithm calculates the fitness values for both parents and children (offspring). Hence, the computational time for every generation is two times

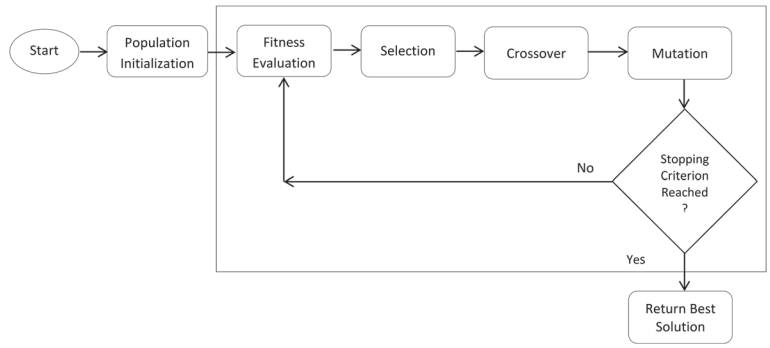

FIGURE 3 Steps in genetic algorithm based feature selection

TABLE 5 Genetic algorithm based feature selection

Input: The training data set with N attributes and K examples

Output: Feature subset

Initialize Parameters

Initialize Population

while (the number of generations)

do

Randomly select two individuals at a time from the population and perform crossover and mutation.

Calculate the fitness of two parents and two children.

Based on the fitness value select the top two individuals to form the new population for the next generation.

end while

longer than the time taken by particle swarm optimization. To overcome this disadvantage, a greedy search based GA for feature selection is proposed. The steps in the proposed algorithm are shown in Figure 4 and summarized in Table 6.

3.3.4 | Fitness evaluation

To evaluate the selected features using a feature selection algorithm, two classifiers are used, namely a decision tree and logistic regression. 10-fold cross-validation is performed with each of these classifiers to calculate the fitness value for the generation. In 10-fold cross-validation, the dataset is divided into 10 parts, nine parts are used for training the model and one part is used for testing the model 10 times. Finally, an average of 10 iterations is considered as the fitness value for the generation.

3.4 | Multimodal fusion

The extraction of contextual information among the neighboring utterances of multimodal data and multimodal fusion are the challenges of multimodal sentiment analysis and affective computing. These steps are discussed in detail in this section.

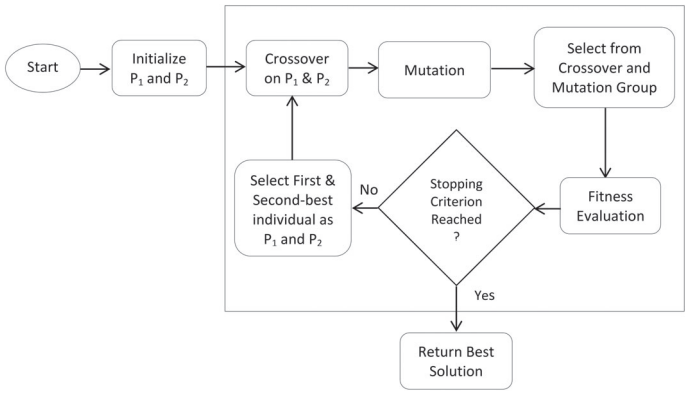

FIGURE 4 Steps in greedy search based genetic algorithm for feature selection

TABLE 6 Greedy search based genetic algorithm for feature selection

Input: *The training data set with N attributes and K examples*

Output: *Feature subset*

Initialize two parents, first parent with all selected and second parent with all unselected.

P1 = [1, 1, 1, … … , 1]

P2 = [0, 0, 0, … … , 0]

while (the number of generations)

do

Perform crossover on two parents P1 and P2 to generate the cross-over group.

Randomly select two individuals from the crossover group and perform mutation.

Each time perform mutation on one bit from the randomly selected individuals and save the result as mutation group.

Select an equal proportion of individuals from the mutation and crossover group.

Calculate the fitness value on the selected population and save the best and second-best genes depending on the fitness value.

The first and second-best individuals are the two parents P1 and P2 for the next generation, respectively.

end while

3.4.1 | Unimodal feature vector

Once the features are extracted from audio, textual and video modalities, a subset of features is selected using a GA, PSO, and GGA. The unimodal feature vector with a subset of features for audio, textual and video modalities is denoted by T_{fs}, $A_{fs,}$ and V_{fs} and represented as:

$$T_{fs} =< T_1, T_2, T_3, \dots \dots, T_n >, \tag{7}$$

$$A_{fs} =< A_1, A_2, A_3, \dots, A_n >, \tag{8}$$

$$V_{fs} =< V_1, V_2, V_3, \ldots, V_n >, \tag{9}$$

where n is the number of utterances and T_i, $A_{i,}$ and V_i are the subset of features selected for i^{th} utterance.

As the video is divided into small utterances as such they may not be complete sentences and the effective opinion or emotion of an utterance may depend on the results of neighboring utterances. Hence, after selecting the important subset of features, bidirectional LSTM based model is used to extract the contextual information among the utterances and is represents as:

$$T_{fsc} = \text{biLSTM}(T_{fs}). \tag{10}$$

$$A_{fsc} = \text{biLSTM}(A_{fs}). \tag{11}$$

$$V_{fsc} = \text{biLSTM}(V_{fs}). \tag{12}$$

3.4.2 | Bimodal feature vector

The bimodal contextual feature vector is constructed by concatenating two of the unimodal contextual feature vectors at a time. The bimodal feature vector with a subset of features is represented as,

$$TA_{fs} =< TA_1, TA_2, TA_3, \ldots, TA_n >, \tag{13}$$

$$TV_{fs} =< TV_1, TV_2, TV_3, \ldots, TV_n >, \tag{14}$$

$$AV_{fs} =< AV_1, AV_2, AV_3, \ldots, AV_n >, \tag{15}$$

where XY_i is the i^{th} utterance bimodal feature vector obtained by concatenating unimodal contextual feature vector with a subset of features of X and Y modalities and is represented by,

$$XY_i = \text{Concatenate}(X_i, Y_i). \tag{16}$$

Furthermore, as in the case of unimodal feature vector construction, the biLSTM based model is used to extract the bimodal contextual information among the utterances and is represented as,

$$TA_{fsc} = \text{biLSTM}(TA_{fs}). \tag{17}$$

$$TV_{fsc} = \text{biLSTM}(TV_{fs}). \tag{18}$$

$$AV_{fsc} = \text{biLSTM}(AV_{fs}). \tag{19}$$

3.4.3 | Trimodal feature vector

The trimodal (audio-text-video) contextual feature vector is formed by concatenating the three modalities. Hence, bimodal contextual feature vectors are fused to get the trimodal contextual feature vector and is represented as,

$$TAV_{fs} =< TAV_1, TAV_2, TAV_3, \dots, TAV_n >, \tag{20}$$

where XYZ_i is the i^{th} utterance trimodal feature vector obtained by concatenating bimodal contextual feature vector with a subset of features of X, Y, and Z modalities and is represented by,

$$XYZ_i = \text{Concatenate}(X_i, Y_i, Z_i). \tag{21}$$

Similar to unimodal and bimodal feature vector construction, the biLSTM based model is used to extract the trimodal contextual information among the utterances and is represented as,

$$TAV_{fsc} = \text{biLSTM}(TAV_{fs}). \tag{22}$$

3.5 | Weighted feature ensemble

The unimodal subset of features is selected using evolutionary computing based feature selection algorithms such as GA, PSO, and GGA. The weighted ensemble feature vector is constructed in three steps. The first union of a unimodal feature vector that is a feature pool is formed. Second, the frequency of each feature is calculated. Finally, the weighted feature vector is constructed using the feature pool and weight of each time.

The union of feature vector is represented as,

$$f_u =< f_1, f_2, f_3, \dots \dots, f_n >, \tag{23}$$

where n is the number of features in the union of features and $f \in \{T, A, V\}$. The weighted ensemble feature vector is represented as,

$$f_w = \sum_{i=0}^{n} w_i * f_i, \tag{24}$$

where w_i is the weight of each feature in the feature pool.

3.6 | Training and classification

The trimodal contextual feature vector with a subset of features TAV_{fsc} is fed as input to a softmax classifier. The softmax classifier predicts the label $\hat{y}$ for testing utterance. The softmax output classifier is represented as,

$$p(y|U) = \text{softmax}\,(\, w_s\, TAV_{fsc} + b_s\,), \tag{25}$$

where w_s and b_s are weight matrix and bias matrix, respectively.

TABLE 7 Proposed multi-level feature optimization and multimodal contextual fusion

1 : *Procedure FeatureExtraction*(U)	*Procedure to extract unimodal features*
2 : *for i in* 1 *to N do*:	
3 : $T_i \leftarrow textualFeatures(U_i)$	
4 : $A_i \leftarrow audioFeatures(U_i)$	
5 : $V_i \leftarrow videoFeatures(U_i)$	
6 : *Procedure FeatureSelection*(F)	*Procedure to extract unimodal subset of features* $fSelect \in \{PSO, GA, GGA\}$
7 : $F_{fs} = fSelect(F)$	
8: *returen* (F_{fs})	
9 : *Procedure ContextExtraction*(F)	*Procedure to extract unimodal context*
10 : $F_{fsc} \leftarrow biLSTM(F)$	
11 : *return* (F_{fsc})	
12 : *Procedure BimodalFusion*(X, Y)	*Procedure for Bimodal fusion where* $X\,! = Y \in \{T, A, V\}$
13 : *for i in* 1 *to N do*:	
14 : $XY_i = Cocatanate(X_i, Y_i)$	
15 : $XY \leftarrow (XY_1, XY_2, \ldots\ldots, XY_N)$	
16 : *return*(XY)	
17: *Procedure TrimodalFusion*(X, Y, Z)	*Procedure for Trimodal fusion where* $X\,! = Y\,! = Z \in \{T, A, V\}$
18 : *for i in* 1 *to N do*:	
19: $XYZ_i = Concatenate(XY_i, XZ_i, YZ_i)$	
20 : $XYZ \leftarrow (XYZ_1, XYZ_2, \ldots, XYZ_N)$	
21 : *return* (XYZ)	
22 : *Procedure Classification* (F)	*Procedure for classification of utterance into discrete number of classes*
23 : $p(y\|F) = softmax\,(\,w_s\,F + b_s\,)$	
24 : $\hat{y} = \arg\max_y\ p\,(y\|F)$	
25 : *return* $(\hat{y})$	
26 : *FeatureExtraction*(U)	*Unimodal Feature Extraction*
27 : *for* $F\,\varepsilon\,\{T, A, V\}$	
28 : $F_{fs} = FeatureSelection(F)$	*Feature Selection*
29 : *for* $F\,\varepsilon\,\{T, A, V\}$	
30 : $F_{fsc} \leftarrow ContextExtraction(F_{fs})$	*Unimodal Context Extraction*
31 : $TA_{fs} \leftarrow BimodalFusion(\,T_{fsc}, A_{fsc})$	*Bimodal Fusion*
32 : $TV_{fs} \leftarrow BimodalFusion(\,T_{fsc}, V_{fsc})$	
33 : $AV_{fs} \leftarrow BimodalFusion(\,A_{fsc}, V_{fsc})$	
34 : *for* $F\,\varepsilon\,\{TA, TV, AV\}$	*Bimodal Context Extraction*
35 : $F_{fsc} \leftarrow ContextExtraction(F_{fs})$	
36 : $TAV_{fs} \leftarrow TrimodalFusion(TA_{fs}, TV_{fs}, AV_{fs})$	*Trimodal Fusion*
37 : $TAV_{fsc} \leftarrow ContextExtraction(TAV_{fs})$	*Trimodal Context Extraction*
38 : *for i in* 1 *to N do*:	*Classification*
39 : $C \leftarrow Classification\,(TAV_{fsc})$	

Let p is the probability of predicted distribution for the utterance classes, that is, happiness, anger, neutral and sadness for affective classification and positive and negative polarity for sentiment classification, then the predicted label $\hat{y}$ for the utterance is defined as,

$$\hat{y} = \arg\max_{y} (p (y|U)). \tag{26}$$

Given the multimodal data, the proposed model is tested by using the cross-entropy loss function $L(\theta)$. The function (cross-entropy) is defined as,

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_i^j \log \hat{y}_i^j + \lambda \sum_{k=0}^{N} \theta_k^2, \tag{27}$$

where N is the number of utterances and M is a number of classes in the training data. y_t is the true label and $\hat{y}_t$ is the predicted label of the t^{th} utterance (the label is either positive or negative for sentiment classification and happiness, anger, sadness or neutral for emotion classification). λ and θ are L2-regularization and parameter set constants, respectively.

The steps in proposed multi-level feature optimization and multimodal contextual fusion are summarized in Table 7.

4 | RESULT ANALYSIS AND DISCUSSION

4.1 | Experimental setup

The feature selection methods such as GA, PSO, and greedy search based GA are implemented in python. The logistic regression and decision tree algorithm from Scikit-Learn Library is used as a fitness function. Also, the multimodal fusion and classification are implemented in python using Keras library with tensor-flow as backend. Experiments are conducted on Tesla K80 GPU with 12GB RAM.

4.2 | Feature selection

Experiments are conducted on two publically available datasets such as IEMOCAP and CMU-MOSI. Initially, accuracy is calculated using all features with decision tree and logistic regression classifiers. The results are shown in Table 8. The textual modality obtains the maximum unimodal accuracy of 77.23% and 77.29% for IEMOCAP and CMU-MOSI dataset with logistic regression as a classifier. The audio modality obtains the minimum accuracy of 55.69% and 65.62% for IEMOCAP and CMU-MOSI dataset with decision tree and logistic regression classifier, respectively.

Next, a subset of features is selected using PSO, GA and GGA based features selection methods. Decision tree and logistic regression classifiers are used as a fitness function. The experimental results on IEMOCAP and CMU-MOSI datasets are shown in Table 9. Population and generation per iteration are set to (30, 100). The results show that a subset of features selected using evolutionary computing based feature selection methods performs better than the unimodal models with all features in terms of classification accuracy. The maximum accuracy

TABLE 8 Unimodal accuracy with all features

	IEMOCAP dataset		CMU-MOSI dataset	
Modality	**Decision tree**	**Logistic regression**	**Decision tree**	**Logistic regression**
Text	72.22	77.23	66.07	77.29
Audio	55.69	60.76	66.85	65.62
Video	68.02	60.88	68.03	72.94

TABLE 9 Subset of features selected using evolutionary computing algorithms

		IEMOCAP dataset				CMU-MOSI dataset			
		Decision tree		**Logistic regression**		**Decision tree**		**Logistic regression**	
	Modality	**NFS**	**Acc**	**NFS**	**Acc**	**NFS**	**Acc**	**NFS**	**Acc**
PSO	Text	53	70.76	51	77.14	31	74.14	58	77.27
	Audio	22	58.47	27	61.52	22	68.73	26	66.55
	Video	49	68.39	25	62.76	29	70.41	52	73.33
GA	Text	53	72.72	61	77.22	43	73.76	47	77.36
	Audio	52	63.04	45	61.67	47	68.73	45	66.55
	Video	**50**	**69.49**	58	62.90	48	70.60	**42**	**73.46**
GGA	Text	96	73.53	**92**	**77.28**	29	74.31	**73**	**77.34**
	Audio	**04**	**68.28**	57	61.69	**32**	**68.86**	18	67.09
	Video	11	69.31	52	62.80	42	70.93	44	73.30

Abbreviations: Acc, accuracy; NFS, no of features selected.

obtained by unimodal modalities is shown with bold letters. GGA based feature selection method achieves maximum accuracy for textual and audio modalities and GA based approach for video modality.

4.3 | Multimodal contextual fusion

The results of proposed multimodal contextual fusion with a subset of features are compared with state-of-the-art baselines.[4,27] CNN based model, OpenSmile toolkit and CLM-Z model are used to extract textual, audio, and visual features, respectively.[4] They construct bimodal and trimodal feature vectors by concatenating the unimodal feature vectors. Multiple-kernel learning classifier was used to classify the bimodal and trimodal feature vectors. Novel fusion technique was proposed in Reference 27 called tensor fusion.

Tables 10 and 11 gives the comparison of results of proposed models with standard baselines on IEMOCAP and CMU-MOSI dataset, respectively. Tables 9-11 show that the bimodal and trimodal models outperform the unimodal model by a huge margin in terms of classification accuracy. Weighted feature ensemble approach achieves the maximum classification accuracy of 83.11% and 82.42% with text-audio and text-audio-video combination of modalities for IEMOCAP and CMU-MOSI dataset, respectively.

TABLE 10 Comparison of results of proposed models on IEMOCAP dataset

			Proposed bi-LSTM models (Accuracy)							
			All features	GA based FS		PSO based FS		GGA base FS		
Modality	Poria et al[4]	Zadeh et al[27]		LR	DT	LR	DT	LR	DT	WFE
T + A	73.7	71.1	77.06	81.04	81.04	81.95	79.55	77.81	80.29	80.71
T + V	74.1	73.7	76.65	81.37	79.22	78.06	78.47	81.70	82.53	83.11
A + V	68.4	67.4	70.00	74.32	72.34	71.68	71.68	72.16	75.00	75.14
T + A + V	74.1	73.6	80.38	81.29	82.20	79.30	80.62	80.79	80.38	82.69

Abbreviations: A, audio; DT, decision tree; FS, feature selection; LR, logistic regression; T, text; V, video; WFE, weighted feature ensemble.

TABLE 11 Comparison of results of proposed models on CMU-MOSI dataset

			Proposed bi-LSTM models (Accuracy)							
			All features	GA based FS		PSO based FS		GGA base FS		
Modality	Poria et al[4]	Zadeh et al[27]		LR	DT	LR	DT	LR	DT	WFE
T + A	77.3	77.0	78.85	80.31	78.98	78.98	79.52	78.85	78.85	80.05
T + V	77.8	77.1	79.12	79.65	79.12	79.38	80.18	79.65	79.12	79.38
A + V	57.9	56.5	65.44	68.06	68.97	65.75	68.65	64.83	68.96	69.54
T + A + V	78.7	77.2	79.52	80.84	80.70	81.22	81.63	80.94	81.49	82.42

TABLE 12 Comparison of the execution time of proposed models for IEMOCAP dataset

	Proposed bi-LSTM based (time in seconds)							
		GA based FS		PSO based FS		GGA Base FS		
Modality	All features	LR	DT	LR	DT	LR	DT	Weighted feature ensemble
T + A	108.47	77.83	76.51	83.47	90.65	82.01	81.91	89.26
T + V	106.48	78.50	78.49	82.54	89.90	80.75	82.32	90.61
A + V	103.63	77.85	79.00	83.48	93.67	84.71	82.13	91.56
T + A + V	135.03	103.94	105.9	88.77	104.16	87.80	88.32	97.73

Tables 12 and 13 shows the execution time of the proposed model with all features and subset of features selected using evolutionary computing algorithms on IEMOCAP and CMU-MOSI datasets, respectively. From the results, it can be observed that the execution time of the proposed models with a subset of features is less compared to all features.

Figures 5 and 6 shows the comparison of experimental results on the IEMOCAP dataset for emotion classification and CMU-MOSI dataset for sentiment classification.

TABLE 13 Comparison of the execution time of proposed models for CMU-MOSI dataset

	Proposed bi-LSTM based (time in seconds)							
		GA based FS		PSO based FS		GGA base FS		
Modality	All features	LR	DT	LR	DT	LR	DT	Weighted feature ensemble
T + A	35.39	29.99	31.71	30.82	30.77	29.72	29.82	30.00
T + V	35.22	29.29	31.59	30.10	30.98	30.50	29.84	30.11
A + V	37.30	30.89	32.38	30.22	31.21	30.89	30.53	31.33
T + A + V	39.46	32.83	34.30	32.42	32.83	31.71	31.39	33.01

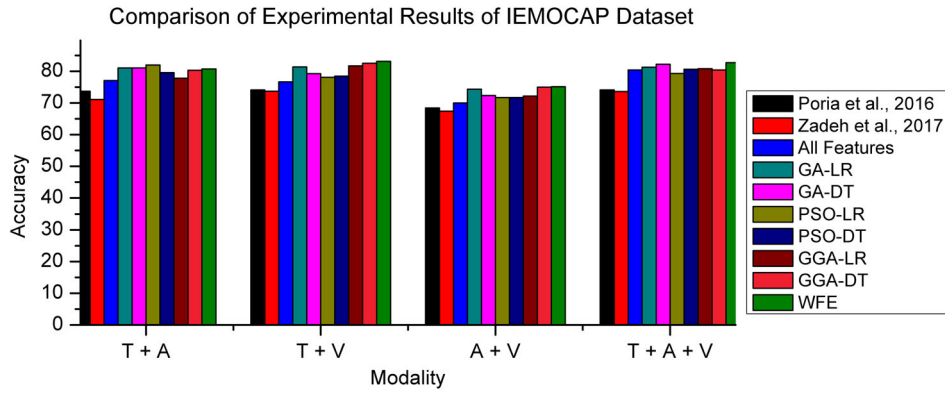

FIGURE 5 Comparison of experimental results on IEMOCAP dataset [Color figure can be viewed at wileyonlinelibrary.com]

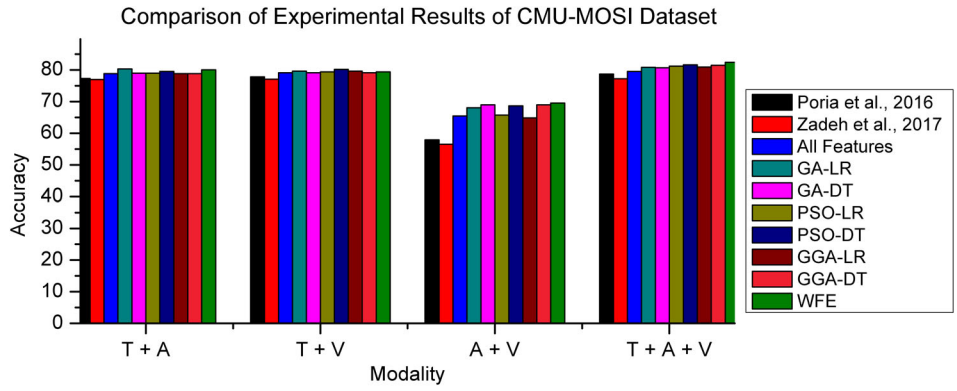

FIGURE 6 Comparison of experimental results on CMU-MOSI dataset [Color figure can be viewed at wileyonlinelibrary.com]

5 | CONCLUSION AND FUTURE WORK

The feature selection, context extraction, and multimodal fusion are the most important challenges in multimodal sentiment analysis and affective computing. The proposed multilevel feature optimization and multimodal contextual fusion method address these issues. First, evolutionary computing based features selection methods are proposed to select an important subset of features from raw unimodal features. The contextual information among the utterance of multimodal data is extracted using the biLSTM model. The bimodal and trimodal feature vector is constructed by concatenating two-two modalities at a time and all modalities, respectively. After every step, a subset of features and contextual information is extracted. Finally, the trimodal contextual feature vector with a subset of features is fed as an input to a softmax classifier. The model is demonstrated on two publically available datasets IEMOCAP and CMU-MOSI. Results show that the classification accuracy of proposed models with a subset of features is higher than the model with all features. Also, the execution time for models with a subset of features is less than a model with all features. In the future, the work can be extended to improve the quality of unimodal features and to select the class-specific features to improve the overall accuracy.

ORCID

Mahesh G. Huddar https://orcid.org/0000-0002-4344-6024
Sanjeev S. Sannakki https://orcid.org/0000-0002-7084-2196
Vijay S. Rajpurohit https://orcid.org/0000-0003-0659-296X

REFERENCES

1. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. *Information Fusion*. 2017a;37:98-125.
2. Huddar MG, Sannakki SS, Rajpurohit VS. Multimodal emotion recognition using facial expressions, body gestures, speech, and text modalities. *Int J Eng Adv Technol*. 2019b;8(5):2453-2459.
3. Pérez-Rosas V, Mihalcea R, Morency L-P. Utterance-level multimodal sentiment analysis. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Bulgaria: Sofia; 2013:973-982.
4. Poria S, Cambria E, Howard N, Huang G-B, Hussain A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*. 2016a;174:50-59.
5. Cambria E. Affective computing and sentiment analysis. *IEEE Intell Syst*. 2016;31(2):102-107.
6. Chaturvedi I, Ragusa E, Gastaldo P, Zunino R, Cambria E. Bayesian network based extreme learning machine for subjectivity detection. *J Frankl Inst*. 2018;355(4):1780-1797.
7. Ma Y, Peng H, Khan T, Cambria E, Hussain A. Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. *Cogn Comput*. 2018;10(4):639-650.
8. Chen L, Zhang H, Jose JM, Yu H, Moshfeghi Y, Triantafillou P. Topic detection and tracking on heterogeneous information. *J Intell Inf Syst*. 2018;51(1):115-137.
9. Jafari M, Shahabi A, Wang J, Qin Y, Tao X, Gheisari M. Automatic text summarization using fuzzy inference. *Proceedings 22nd International Conference on Automation and Computing*. Colchester, England: IEEE Conference; 2016:256-260.
10. Liu B, Zhang L. A survey of opinion mining and sentiment analysis. *Mining Text Data*. Boston, MA: Springer; 2012.
11. Huddar MG, Sannakki SS, Rajpurohit VS. An ensemble approach to utterance level multimodal sentiment analysis. *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. Belgaum, India: IEEE; 2018:145-150.
12. Kiritchenko S, Zhu X, Mohammad SM. Sentiment analysis of short informal texts. *J Artif Intell Res*. 2014;50:723-762.
13. Badaro G, Jundi H, Hajj H, El-Hajj W. EmoWordNet: automatic expansion of emotion lexicon using English WordNet. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. New Orleans, Louisiana: Association for Computational Linguistics; 2018:86-93.

14. Nalisnick ET, Baird HS. Extracting sentiment networks from shakespeare's plays. *12th International Conference on Document Analysis and Recognition*. Washington, DC: IEEE; 2013:758-762.
15. Thakor P, Sasi S. Ontology-based sentiment analysis process for social media content. *Proc Comput Sci*. 2015;53:199-207.
16. Dash M, Liu H. Feature selection for classification. *Intell Data Anal*. 1997;1(1–4):131-156.
17. Liu H, Zhao Z. Manipulating data and dimension reduction methods: feature selection. *Encyclopedia of Complexity and Systems Science*. Berlin, Germany: Springer; 2009:5348-5359.
18. Whitney AW. A direct method of nonparametric measurement selection. *IEEE Trans Comput*. 1971;100(9):1100-1103.
19. Min F, Hu Q, Zhu W. Feature selection with test cost constraint. *Int J Approx Reason*. 2014;55(1):167-179.
20. Huddar MG, Sannakki SS, Rajpurohit VS. A survey of computational approaches and challenges in multimodal sentiment analysis. *Int J Comput Sci Eng*. 2019a;7(1):876-883.
21. Wöllmer M, Weninger F, Knaup T, et al. Youtube movie reviews: sentiment analysis in an audio-visual context. *IEEE Intell Syst*. 2013;28(3):46-53.
22. Rosas VP, Mihalcea R, Morency L-P. Multimodal sentiment analysis of Spanish online videos. *IEEE Intell Syst*. 2013;28(3):38-45.
23. Zadeh A. Micro-opinion sentiment intensity analysis and summarization in online videos. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. Seattle, Washington: Association for Computing Machinery; 2015:587-591.
24. Poria S, Chaturvedi I, Cambria E, Hussain A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. *IEEE 16th International Conference on Data Mining (ICDM)*. Barcelona, Spain: IEEE; 2016b:439-448.
25. Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl-Based Syst*. 2018;161:124-133.
26. Pham H, Liang PP, Manzini T, Morency L-P, Póczos B. Seq2seq2sentiment:Multimodal sequence to sequence models for sentiment analysis. *Proceedings of the First Grand Challenge and Workshop onHuman Multimodal Language (Challenge-HML)*. Melbourne, Australia: Association for Computational Linguistics; 2018:53-63.
27. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P. Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics; 2017:1103-1114.
28. Busso C, Bulut M, Lee C-C, et al. IEMOCAP: interactive emotional dyadic motion capture database. *J Lang Resour Eval*. 2008;42(4):335-359.
29. Zadeh A, Zellers R, Pincus E, Morency L-P. Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. *IEEE Intell Syst*. 2016;31(6):82-88.
30. Karpathy A, Toderici G, Toderici G, et al. Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus,OH: IEEE; 2014.
31. Teh YW, Hinton GE. Rate-coded restricted Boltzmann machines for face recognition. *Advances in Neural Information Processing Systems*. Cambridge, MA; 2001.
32. Florian Eyben FWFGBS. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. *Proceedings of the 21st ACM International Conference on Multimedia*. Barcelona, Spain: ACM; 2013:835-838.
33. Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. *IEEE Trans Pattern Anal Mach Intell*. 2012;35(1):221-231.
34. Yang H, Yuan C, Li B, et al. Asymmetric 3D convolutional neural networks for action recognition. *Pattern Recogn*. 2019;85:1-12.
35. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P. Context-dependent sentiment analysis in user-generated. *ACL*. 2017b;2:873-883.

How to cite this article: Huddar MG, Sannakki SS, Rajpurohit VS. Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification. *Computational Intelligence*. 2020;36:861–881. https://doi.org/10.1111/coin.12274