Accepted refereed manuscript of: Dashtipour K, Gogate M, Cambria E & Hussain A (2021) A Novel Context-Aware Multimodal Framework for Persian Sentiment Analysis. *Neurocomputing*, 457, pp. 377-388.
https://doi.org/10.1016/j.neucom.2021.02.020

A Novel Context-Aware Multimodal Framework for Persian Sentiment Analysis

Kia Dashtipour[a], Mandar Gogate[b], Erik Cambria[c], Amir Hussain[b]

[a] *Department of Computing Science and Mathematics, University of Stirling, Stirling, UK*
[b] *School of Computing, Edinburgh Napier University, Edinburgh, UK*
[c] *School of Computer Science and Engineering, Nanyang Technological University, Singapore*

Abstract

Most recent works on sentiment analysis have exploited the text modality. However, millions of hours of video recordings posted on social media platforms everyday hold vital unstructured information that can be exploited to more effectively gauge public perception. Multimodal sentiment analysis offers an innovative solution to computationally understand and harvest sentiments from videos by contextually exploiting audio, visual and textual cues. In this paper, we, firstly, present a first of its kind Persian multimodal dataset comprising more than 800 utterances, as a benchmark resource for researchers to evaluate multimodal sentiment analysis approaches in Persian language. Secondly, we present a novel context-aware multimodal sentiment analysis framework, that simultaneously exploits acoustic, visual and textual cues to more accurately determine the expressed sentiment. We employ both decision-level (late) and feature-level (early) fusion methods to integrate affective cross-modal information. Experimental results demonstrate that the contextual integration of multimodal features such as textual, acoustic and visual features deliver better performance (91.39%) compared to unimodal features (89.24%).

Keywords: `Multimodal Sentiment Analysis, Persian Sentiment Analysis`

*Corresponding author: Mandar Gogate
Email address: `M.Gogate@napier.ac.uk` (Mandar Gogate)

1. Introduction

The emergence of online forums has increased exponentially in the last few years [1]. Furthermore, with the advent of social media (e.g., Twitter, YouTube and Facebook), people can share their opinions frequently. Social media encourage people to engage in discussions and enables them to share their thoughts on a range of issues and challenges [2]. Online media provide a platform for sharing ideas and encourages public to join group discussions. In addition, social media allow companies and organizations to get feedback regarding their products in the form of texts, images and videos [3, 4, 3, 5, 6, 7, 8].

To date, most current research in sentiment analysis has been conducted on textual data [9, 10, 11, 12]. As a result, most developed resources are limited to sentiment analysis from text. However, with advent of social media, people are extensively using social media to express their opinions in different languages [13, 14, 15]. In addition, they are increasingly using videos, audio, images and text to express their opinions in social media. Therefore, it is becoming vital to identify sentiment in various modalities as well as in different languages [16, 17]. An overview of a typical multimodal sentiment analysis framework is shown in Fig. 1, which follows our pioneering works in [18, 19, 20, 21, 22, 23]. As shown in Fig. 1, the visual, acoustic and textual features are extracted to determines the overall polarity of the input videos. First, the video is transcribed into text, then, the acoustic features are extracted using openSMILE and finally visual features extracted and all features concatenated to find overall polarity of the sentence.

Persian comprises 32 letters, which cover 28 Arabic letters. Its writing system includes special signs and diacritic marks that can be used in different forms or omitted from the word. The Persian language has more than 100 million speakers in Iran, Afghanistan and Tajikistan. Key challenges in Persian language are sarcasm, idioms, use of informal words, phrases and more spelling mistakes compared to other languages such as English [25, 26, 27]. In our recent work [28], we proposed a first of its kind, linguistic dependency-rules (pattern)

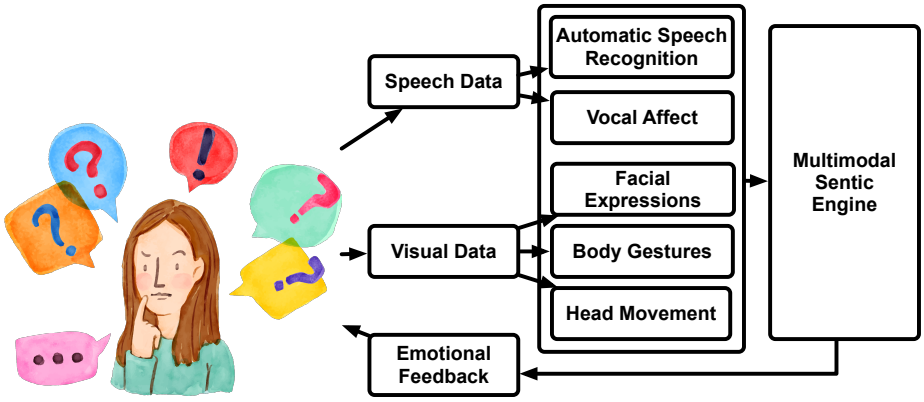

Fig. 1: A generic multimodal sentiment analysis framework [24]

based deep learning approach for Persian sentiment analysis.

In this paper, we address the more challenging task of Persian multimodal sentiment analysis, by developing a novel context-aware framework and conduct an extensive set of experiments to show that a multimodal model that contextually combines visual, acoustic and text features can significantly enhance the sentiment polarity detection of online Persian videos.

The main aim of this paper is to address the task of multimodal sentiment analysis by conducting a proof-of-concept set of experiments to show that the multimodal features including visual, acoustic and textual features in Persian language can be more effective to identify the overall sentiment of multimodal data. Therefore, in this paper, we address the first-time task for tri-modal sentiment analysis by integrating three different modalities (audio, visual and text) which are integrated to identify the polarity of the input data. This is unlike the previous studies on sentiment analysis which only considers text modality. Second, we presented a first of its kind novel Persian multimodal sentiment analysis data collected from YouTube to enable training of multimodal sentiment analysis models. The experimental results show that the multimodal features (textual, acoustic and visual) achieved better performance as compared unimodal features.

In summary, the paper reports four major contributions outlined below:

1. We propose a novel context-aware multimodal fusion framework to extract unimodal and multimodal features using automated feature extraction methods including convolutional neural network (CNN) and long short-term memory (LSTM). The features are integrated to perform multimodal sentiment analysis task.
2. A first of its kind Persian multimodal dataset is presented, consisting of video opinions, collected from YouTube, which is analyzed and annotated to enable multimodal sentiment analysis.
3. We conduct an extensive set of experiments to demonstrate that a multimodal approach contextually exploiting visual, acoustic and textual features can more accurately determine the sentiment polarity of online Persian videos, compared to state-of-the-art unimodal approaches (e.g. text-based).
4. The context-aware multimodal approach is shown to overcome the limitations of ambiguous words usage in Persian. For example, Persian textual reviews often comprise ambiguous words and phrases such as فیلم افتضاح خوب بود, 'The movie is awfully good' which can lead to incorrect classification of data. In such cases, the contextual use of acoustic and visual features enable more accurate polarity assignment through the different modalities.

The rest of the paper is organized as follows: Section 2 presents related work on state-of-the-art multimodal sentiment analysis approaches for English and other languages; Section 3 presents the proposed framework for Persian multimodal sentiment analysis; Section 4 presents the novel Persian multimodal dataset; Section 5 presents comparative experimental results; finally, Section 6

concludes this paper and outlines some future work directions.

2. Related Work

In the literature, extensive research has been carried out to develop multimodal models, that demonstrate the significance of multimodal processing over unimodal approaches. Researchers have proposed multimodal processing architectures for a wide range of real-world applications ranging from sentiment analysis [29, 30, 31, 32, 33, 34, 35, 36], deception detection [37] to dementia diagnosis and progression prediction [38] and audio-visual speech recognition and enhancement [39, 40, 41, 42].

In related work, Morency [43] presented an approach for Spanish multimodal sentiment analysis using the MOUD dataset which is product reviews videos collected from YouTube to evaluate its performance. Comparative experimental results showed that the combination of audio, visual and text features can improve the performance of the approach. Poria et al. [44] presented a novel method to extract text features for short text as part of a multimodal sentiment analysis framework. Their dataset comprised 498 (in English) short videos and fused audio, text and visual features were used to train their model.

Poria et al. [24] also proposed a method for English multimodal sentiment analysis to identify sentiments using different approaches to fuse audio, text and video features. Their proposed method was evaluated with a YouTube dataset, however it was restricted to a single speaker and was unable to identify the overall polarity of video with multiple speakers. On the other hand, Rosas et al. [45] proposed an approach for Spanish multimodal sentiment analysis. Due to lack of resources, Spanish videos were collected from YouTube to enable audio, video and textual features extraction. A support vector machine (SVM) was used to evaluate the performance of the approach. Experimental results showed that the combination of text, audio and video features achieved better performance as compared to text and audio modalities. Their approach identified the positive or negative polarity of sentence but was unable to detect the polarity for

neutral sentences.

Alqarafi et al. [46] proposed a multimodal approach to detect polarity in Arabic videos. A total of 40 videos in Arabic language were collected from YouTube and manually transcribed. It is worth to mention that, the collected videos is only cover Arabic Saudi accent and it cannot generalized on different Arabic dialects. In order to evaluate the performance of their approach, text features such as ngram and visual features (smile, frown, head nod, and head shake) were extracted. Empirical results demonstrated that the combination of visual and text features was more effective as compared to unimodal features. Their approach did not however consider acoustic features and only relied on text and visual features.

Dastgheib et al. [47] proposed a novel hybrid method using a combination of feature-based and deep convolutional neural network (CNN) to detect polarity in Persian movie reviews. The results show the proposed hybrid method displayed the combination of feature-based and CNN achieved better performance as compared to CNN and LSTM. Farahani et al. [48] used a BERT architecture to develop a pre-trained framework to detect polarity in Persian product reviews, which is called ParsBERT. The model is based on different NLP tasks such as named entity recognition and sentiment analysis. The experimental results show that the proposed architecture achieved better performance as compared to multilingual approaches.

However, none of the aforementioned works have explored polarity detection challenges for Persian multimodal sentiment analysis. Motivated by this, in this paper we propose a novel framework coupled with a first of its kind multimodal Persian dataset, which are described in Section 4. Table 1 shows the summary of the proposed approach in multimodal sentiment analysis.

3. Methodology

In this section, the proposed multimodal Persian sentiment analysis framework, shown in Fig. 2, is described. As can be seen, audio, visual and textual

Table 1: Summary of the multimodal sentiment analysis approaches

Ref	**Method**	**Dataset**	**Accuracy**
Alqarafi et al. [46]	Arabic Multimodal	YouTube Videoes Arabic Saudi dialects	65.59
Dashtipour et al. [28]	Hybrid Rule-based	Persian movie review Only text modality	86.29
Dastgheib et al. [47]	Hybrid Approach	Persian product reviews Only text modality	74
Farahani et al. [48]	BERT Method	Persian product reviews Only textual modality	98.79
Poria et al.[24]	Multimodal SA	English YouTube videos	77.01
Rosas et al [45]	Multimodal SA	YouTube videos Spanish	75

features are first contextually extracted, and the extracted features, representing affective information, are then fused to identify the overall polarity of the target video dataset.

3.1. Unimodal Feature Extraction

In this section, we discuss unimodal feature extraction techniques.

3.1.1. Textual Feature Extraction: text-BiLSTM

For contextually extracting features from the textual modality, a stacked bidirectional LSTM (BiLSTM) model, as shown in Fig. **??** is used. Each utterance is represented as a concatenation of 300 dimensional pretrained fastText word embeddings. Each utterance is either trimmed with a window of size 60 words or zero padded at the end to form a vector of dimension 60 x 300. The converted vectors are fed into a stacked BiLSTM model, whose parameters are experimentally determined and optimized using a trial and error approach. The model consists of two bidirectional LSTM layers with 128 cells each. The output of the last bidirectional LSTM is concatenated and fed to a fully connected layer with 128 neurons (ReLU activation) and 2 neurons (Softmax activation)

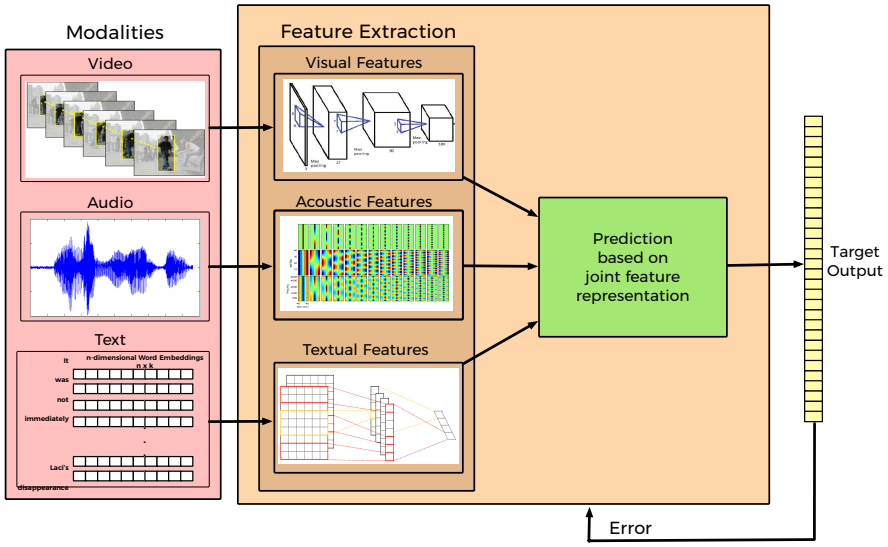

Fig. 2: Overview of the Persian Multimodal Sentiment Analysis Framework

respectively. The network learn the levels of abstract representations and implicit semantic information, that spans over the entire utterance. The proposed implemented BiLSTM architecture is similar to the one which is used by Wang et al. [49], contains of an input layer, two stacked LSTM layers along with an output fully connected layer. Particularly, the BiLSTM consists of two stacked bidirectional LSTM with 128 and 64 cells and dropout of 0.2 probability and dense layer with two neurons and softmax activation.

Following our recent work, we exploit a novel linguistic dependency rules (patterns) and deep learning based approach [28], for the transcribed Persian utterances. As can be seen from example results in Table 2, a deep learning based classifier can be used to effectively classify the unclassified Persian sentences, specifically sentences that cannot be categorized by dependency-based rules. For the results shown in Table 2, each sentence is converted into a 300-dimensional vector using fastText and the concatenation of word embedding is fed into deep learning classifiers. For comparison, the sentences are converted

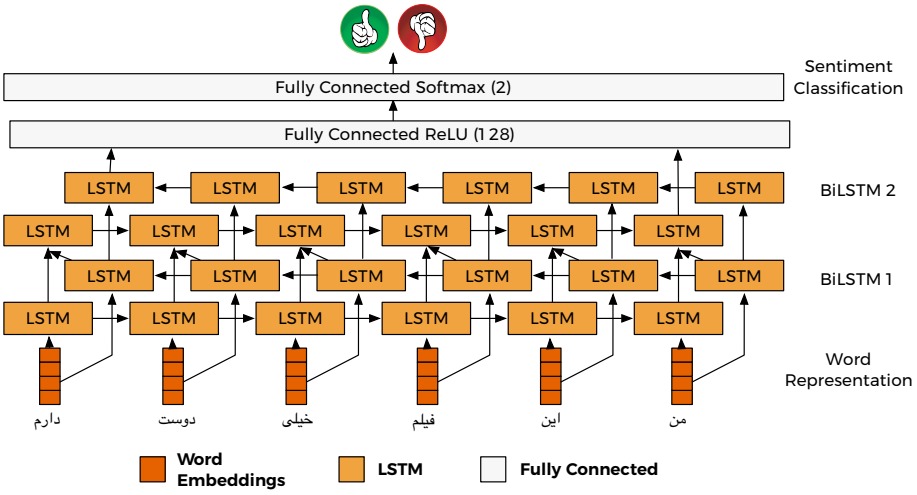

Fig. 3: Bidirectional LSTM architecture for Textual-cues feature extraction

into Bag-of-words and fed into logistic regression and SVM. In addition, the transcribed videos are converted into 300 dimensional word embeddings [50], which are then fed into deep learning (CNN and LSTM) classifiers, with results shown in the Table 2.

The dataset used for the experiments divided into 60% train set, 10% validation set and 30% testing set. The CNN, LSTM, SVM and LR were trained on the train set, tuned on validation set and evaluated on the test set. Comparative experimental results show that the hybrid 2: LSTM + Dependency-based achieved better accuracy as compared with other approaches including DNN based classifiers. The number of neutral discourses are very less as compared to positive and negative cases. Therefore, we cannot train our proposed models on neutral comments. As deep learning requires a large number of training and testing set.

3.2. Audio Feature Extraction: openSMILE

The current literature shows that the openSMILE can effectively extract audio features such as deception as well as sentiment [54, 44]. Therefore, the openSMILE is used because it can automatically extract low-level descriptors

Table 2: Textual Features Results

Classifier	Precision	Recall	F-measure	Accuracy
Kim et al. [51]	0.61	0.63	0.61	61.24
Dehkharghani et al. [52]	0.78	0.82	0.79	70.12
fastText Classifier [53]	0.67	0.67	0.67	70.01
SVM	0.65	0.65	0.65	65.01
Logistic Regression	0.64	0.64	0.64	64.23
Dependency-based Approach	0.83	0.93	0.87	75.94
CNN	0.91	0.63	0.75	68.53
LSTM	0.92	0.83	0.87	86.14
Hybrid 1: CNN + Dependency-Based	0.78	0.77	0.77	76.16
Hybrid 2: LSTM + Dependency-Based	0.86	0.95	0.85	88.01

such as beat histogram, Mel frequency cepstral coefficients, spectral centroid, spectral flux, beat histogram, beat sum.

The audio features are automatically extracted from the speech of each utterance using the widely used OpenSMILE software. The openSMILE is used to extract features which consists of several low level descriptors and their statistical information. In addition, features such as amplitude mean, arithmetic, mean, quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles are also extracted. In total, we have obtained more than 6,373 features. The features are extracted at a frequency of 40 samples per second. The extracted features consist of the following acoustic sub-features:

- Prosody feature: This feature consists of intensity, loudness and pitch that describe the speech signal in terms of amplitude and frequency.
- Energy features: The energy feature depicts the human loudness perception.

Table 3: MLP Architecture for audio feature extraction

Layer	1	2	3	4
Type	Relu	ReLU	ReLU	ReLU
Neurons	1024	512	128	1

- Voice probabilities: The voice probabilities provide an estimate of percentage of voiced and unvoiced energy in the audio.
- Spectral features: The spectral features are based on the characteristics of the human ear, which uses a nonlinear frequency unit to simulate the auditory system
- Cepstral features: The cepstral features emphasize the changes in the spectrum features measured by frequencies. Specifically, 12-model Mel-frequency cepstral coefficients are used and calculated based on the Fourier transform of a speech fame.

The overall audio features for a single utterance consist of 6373 features. Speaker normalization is performed using z-standardization. The voice intensity is a threshold to identify the samples with and without speech. The features are averaged over all the frames in an utterance, to obtain a feature vector for each utterance. The audio feature extraction framework is shown in Fig. 4. The multilayer perceptron (MLP) architecture, as depicted in Table 3, is used to exploit the extracted audio feature for determining the opinion strength based on acoustic cues. The MLP architecture is shown in Fig. 4.

3.3. Visual Feature Extraction: 3D-CNN

Human expressions play a significant role in identifying the emotion expressed in day-to-day conversations [55, 56]. In particular, facial expressions help in decoding the expressed affect by providing visual cues. Therefore, visual features are important in multimodal sentiment analysis. In this work, a Facial Action Coding System (FACS) is used for measuring and describing facial

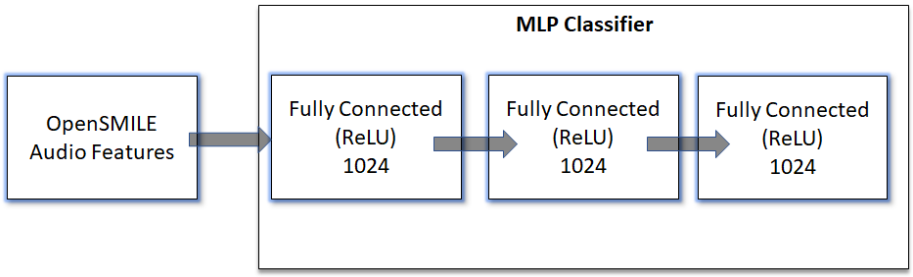

Fig. 4: Acoustic cues feature extraction

behaviors. According to [57], facial behavior can be categorized into 64 action units. The Computer Expression Recognition Toolbox (CERT) is employed to automatically extract the following visual features:

Smile and head pose estimates: The smile feature depicts the probability of a person smiling given an image. Head pose detection consists of three dimensional head orientation yaw, pitch and roll. These dimensions provide information about the face position while expressing positive or negative opinions [58].

Facial action units: The facial action units estimate the thirty related muscle movements related to eyes, nose, eyebrows and chin. This feature provides information about facial behaviors which can be exploited to find differences between positive and negative opinions [59].

The visual features are extracted from videos using a 3D CNN. The 3D-CNN contextually exploits both spatial and temporal patterns to accurately find the spatio-temporal association between a subjective and an objective utterance. In our experiments, the best results are obtained with a 9-layered 3D-CNN architecture as illustrated in Fig. 5. The architecture details are presented in Table 4.

3.4. Multimodal Fusion

In this section, we discuss multimodal fusion techniques. Multimodal fusion is the process of contextually integrating features collected from different

Table 4: 3D-CNN Architecture for visual cues feature extraction)

Layer	**Type**	**Feature Map**	**Kernel**
1	Convolutional3D	16	2 x 2 x 2
2	Convolutional3D	32	2 x 2 x 2
3	Max pooling3D		1 x 2 x 2
4	Convolutional3D	64	2 x 2 x 2
5	Max Pooling3D		2 x 2 x 2
6	convolution3D	64	2 x 2 x 2
7	Max pooling3D		1 x 2 x 2
8	Fully connected	5000	
9	Fully connected	500	
10	Fully connected	2	

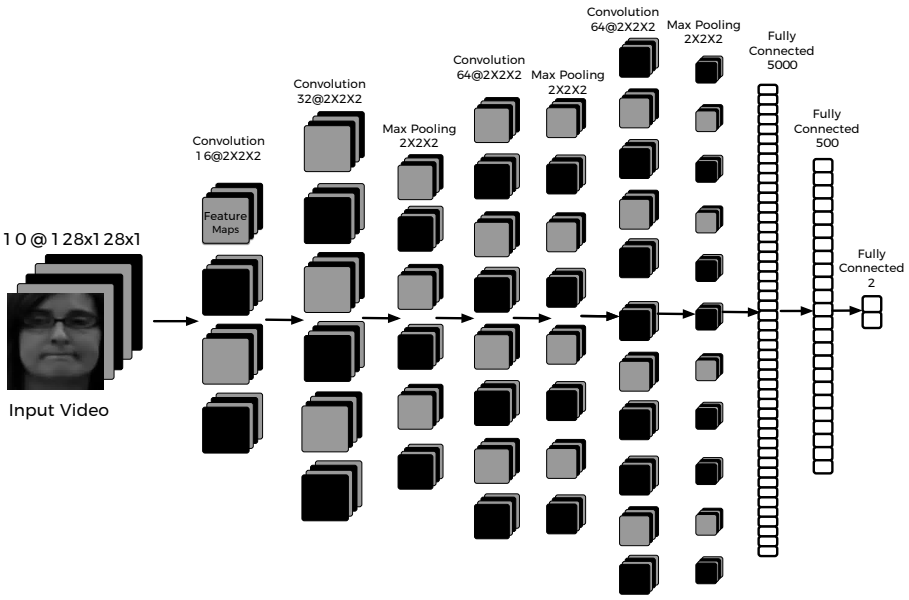

Fig. 5: 3D-CNN based Visual Feature Extraction

modalities for the sentiment analysis task. In general, multimodal fusion techniques can be divided into early or feature-level fusion and late or decision-level

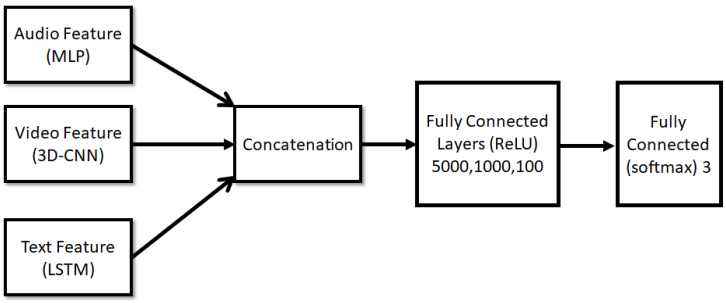

Fig. 6: Early (Feature-level) Fusion

fusion. Multimodal systems often outperform unimodal approaches as the correlation and discrepancies between modalities can help in achieving superior performance [60].

3.4.1. Feature-level (Early) Fusion

In early fusion, first, the features are extracted from input modalities using either deep neural networks or state-of-the-art feature extraction techniques. The input features are concatenated and fed into a classifier. For example, audio features are extracted using OpenSMILE software, visual features are extracted using 3D-CNN and text features are extracted using BiLSTM. The features are then fused and fed into a shallow or deep learning classifier. The main advantage of feature-level fusion is that the cross-correlation between multiple modalities at an early stage helps in achieving better performance. On the other hand, the main disadvantage of early fusion is that the modalities must be tightly time synchronized, since incorrect time synchronization can lead to a poorly functioning system as the model will be unable to learn any cross-modal correlation [61].

3.5. Decision-Level (Late) Fusion

In late or decision-level fusion, unimodal classifiers are used to identify the prediction for each modality. The local predictions are concatenated and further classified to achieve the final decision. The advantage of late fusion is that, the sampling rate at which the predictions are generated is the same, therefore the

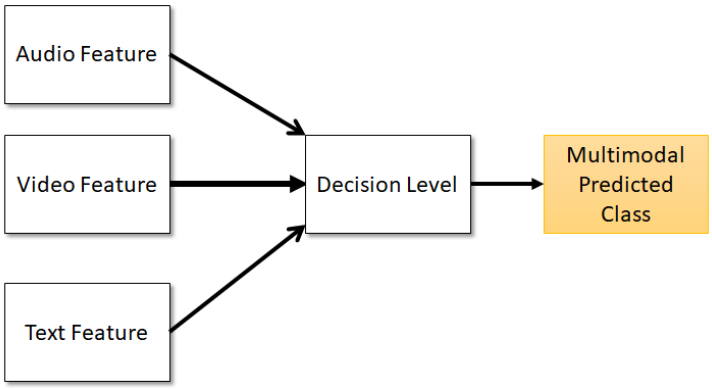

Fig. 7: Late (Decision-Level) Fusion

extracted predictions can be easily concatenated without up-sampling or down-sampling. However, the disadvantage is that the model cannot learn cross-modal correlations [62]. The Fig displays the late fusion-level approach.

4. Persian Multimodal Dataset

In this section, we introduce the novel Persian multimodal dataset.

4.1. Dataset Collection

The YouTube website was used to collect Persian videos with a focus on product, movie and music reviews. The videos were found using keywords such as: نقد فیلم ("Movie critize"), محصولات ایرانی ("Persian product"), نقد موزیك ("Critize music").

A total of 91 videos were selected that respected the following guidelines:

- The video must contain only one speaker
- The speaker should look directly in the camera
- The speaker face should be clearly visible
- There should not be any background noise or music in the recording
- The video should be recorded with high quality microphone and camera

Fig. 8: Example snapshots of videos from our new Persian multimodal dataset

While in the collected videos, the speaker has similar distance from camera, the background and lighting is variable between different videos. The length of the video is between 1-5 minutes. In addition, each video consists of 10–50 utterances. The dataset includes 15 male and 9 female speakers between 20 to 60 years age. Sample snapshots of our Persian multimodal dataset are shown in Fig. 8.

4.2. Segmentation and Transcription

All collected videos were transcribed manually with utterance start and end times. The transcription process comprised two stages. First, an expert transcriber manually transcribed all the videos. The transcribed text was reviewed by two other native speakers. In the second stage, the transcriptions were divided into utterances using pause details such as اه هم (hm, oh, etc.).

Each video was segmented into an average ten utterances, resulting in a final dataset of 945 utterances with 754 subjective utterances and 191 objective utterances. Each utterance was then linked to audio and video streams as well as to manual transcription. The utterances have an average duration of 6 seconds.

Table 5: Persian Multimodal dataset statistics

Total number of positive segmented	468
Total number of negative segmented	366
Total number of subjective	834
Total number of objective	180
Total number of unique words in the dataset	4065
Total number of speakers	24

Three expert annotators categorized segmented utterances into subjective utterances (with expressed emotion) and objective utterances with no polarity (e.g., facts, figures etc.) category. The subjective segmentation was important to achieve fine-grained sentiment analysis. An utterance was categorized into the subjective category, if the sentence was carrying an opinion, belief, thought, feeling or emotion. Three rules were used to identify subjectivity in the sentence as outlined below:

- Explicitly criticising an entity. For example: يك فيلم كمدي طنز خوب كه فراهم ميكنه لهجه هاي خوب بري مخاطب, "The comedy movie is really good and the viewers have good time when they are watching the movie".

- Referencing an opinion expressed by a third person. For example, منتقدين از ديدن فيلم راضي نيستند, "The movie critic are not satisfying with watching this film"

- Implicitly expressing a subjective opinion. من پيشنهاد نميكنم كه اين فيلم با خانواده نگاه كنيد, "I am not recommended to watch this movie with family"

Detailed statistics of the dataset can be found in Table 5.

4.3. Sentiment Polarity

The utterances were annotated by three native Persian speakers (two male and one female) between 30 to 50 years old. All the speakers received a Master

Table 6: Example utterances from the Persian Multimodal Sentiment Analysis Dataset

Persian Sentence	**English Translation**	**Polarity**
بعد چند سال شادمهر يك كار خوب منتشر كرد به اسم با تو عشقم	After few years, Shadmehr release good work called "with you my love	1
بازيگر دوست داشتني رضا عطآران ما رو به ديدن فيلم ترغيب ميكنه	The good acting of Reza Attaran help to people watch the movie	1
ارشاد قصد مجوز به فيلم نداره	The ministry of culture did not give the permission to the movie	-1
تو دوران پهلوي يك مدت به خاطر ترانه سياسي زندان بود	During Pahlavi Dynasty he was jailed for few years	-1

in Persian language. The annotators had three choices, positive (+1), neutral (0) and negative (-1). The polarity assignment included all three modalities (visual, audio and text). Table 6 shows examples of utterances obtained from one of the videos in the multimodal dataset, along with their translation and polarity. It can be observed that, a single video consists of both positive and negative utterances.

Finally, manual gestures including smile, frown, head node and head shake were annotated manually to study the relation between words and gestures. The annotation was carried out by marking the utterances into these gestures. Expert coders manually annotated each utterance with gesture information. The average agreement of the gestures was 89.23%.

Table 7: Prediction Results: Text-Based unimodal Sentiment Analysis

	Precision	**Recall**	**F-measure**	**Accuracy**
Positive	0.92	0.83	0.87	
Negative	0.88	0.94	0.91	
Average	0.90	0.88	0.89	89.24

5. Experimental Results

In this section, comparative experimental results for Persian multimodal dataset are discussed in detail. In this study, we focus on identifying effective strengths as compared to finding if an utterance is subjective or objective. Therefore, we removed all objective utterances for training the multimodal sentiment analysis framework. The features for text, audio and video modalities are experimentally determined and contextually utilized as described in the previous sections.

The results for text-based, audio-based and video-based unimodal sentiment analysis are summarized in Table 7, Table 8 and Table 9 respectively. Experimental results show that the text-based classifiers achieved better accuracy as compared to audio-based and video-based unimodal sentiment analysis models. It is to be noted that the model is trained for utterance level sentiment analysis and not video level sentiment analysis. Generally, each video consists of 10-15 sentences. In addition, it is to be noted that there was no overlap of speakers in training, validation and test data for speaker independent analysis. Therefore, it can be concluded that the model generalizes well on unseen samples.

Table 8 shows the experimental results for audio-based unimodal sentiment analysis. It can be seen that the positive features achieve better recall compared to negative features. However, the negative features achieve better F-measure as compared to positive features.

Table 9 displays the results for video-based unimodal sentiment analysis. As discussed earlier in section 3.3, the 3D-CNN is used to extract features from videos. Experimental results show that the negative features achieve better

Table 8: Prediction Results: Audio-Based unimodal Sentiment Analysis

	Precision	**Recall**	**F-measure**	**Accuracy**
Positive	0.78	0.84	0.81	
Negative	0.78	0.82	0.84	
Average	0.78	0.83	0.82	82.79

Table 9: Prediction Results: Video-Based unimodal Sentiment Analysis

	Precision	**Recall**	**F-measure**	**Accuracy**
Positive	0.76	0.85	0.80	
Negative	0.87	0.79	0.83	
Average	0.81	0.82	0.81	81.72

precision and F-measure as compared to positive features. However, the positive features achieve better recall as compared to negative features. Experimental results show that the visual feature achieved lower performance as compared to acoustic and textual features. The main reason, the speakers have shown less emotion in the visual features as compared to other features.

Table 10 summarizes the results for late or decision-level fusion. As can be seen, comparative experimental results demonstrate that the full multimodal system comprising audio (A) + video (V) + text (T) modalities, and the part-multimodal system comprising V + T modalities achieve better accuracy as compared to other modality combinations. However, the A + T modality achieves better precision compared to other modalities. In addition, the V + T modality achieves better recall compared to the other modalities, whereas the full A + V + T modality achieves better F-measure compared to other modalities. Finally, the results show that the A + V modality achieves the least accuracy as compared to other multimodality combinations. As the experimental results show our proposed method is robust and achieves great performance even if one of the modalities such as visual, acoustic and textual features are not available. For example, the model achieved superior performance using A+V (audio and

Table 10: Prediction Results: Late (decision-level) Fusion

Modality		**Precision**	**Recall**	**F-measure**	**Accuracy**
A+V	Positive	0.78	0.79	0.79	
	Negative	0.84	0.83	0.83	
	Average	0.79	0.78	0.78	81.18
V+T	Positive	0.89	0.89	0.89	
	Negative	0.91	0.91	0.91	
	Average	0.88	0.88	0.88	90.32
A+T	Positive	0.84	0.93	0.88	
	Negative	0.94	0.87	0.90	
	Average	0.92	0.84	0.88	89.24
A+V+T	Positive	0.88	0.90	0.89	
	Negative	0.92	0.90	0.92	
	Average	0.90	0.87	0.89	90.32
T-Only	Average	0.90	0.88	0.89	89.24
A-Only	Average	0.78	0.83	0.82	82.79
V-Only	Average	0.81	0.82	0.81	81.72

visual) without using T (textual). The multimodal systems often exploit the correlation between modalities for predicting more accurate output as compared to unimodal systems. E.g. human facial expressions are often correlated with other modality such as their voice and the words which they are using [63].

Table 11 summarizes the results for early or feature-level fusion. Comparative experimental results show that the A + V + T and V + T modalities achieve better accuracy compared to other modalities. However, the A + T modality achieves better precision compared to other modalities. Furthermore, the V + T modality achieves better recall as compared to other modalities. Finally, the results show that the A + V modality combination achieves the least accuracy compared to other modalities.

Fig. 9 presents the accuracy of unimodal sentiment analysis models for text,

Table 11: Prediction Results: Early (feature-level) Fusion

Modality		**Precision**	**Recall**	**F-measure**	**Accuracy**
A+V	Positive	0.82	0.79	0.81	
	Negative	0.84	0.87	0.85	
	Average	0.79	0.82	0.80	83.33
V+T	Positive	0.90	0.89	0.89	
	Negative	0.92	0.92	0.92	
	Average	088	0.90	0.89	90.86
A+T	Positive	0.91	0.88	0.89	
	Negative	0.91	0.93	0.92	
	Average	0.87	0.91	0.89	90.86
A+V+T	Positive	0.85	0.98	0.91	
	Negative	0.98	0.87	0.92	
	Average	0.97	0.84	0.90	91.39
T-Only	Average	0.90	0.88	0.89	89.24
A-Only	Average	0.78	0.83	0.82	82.79
V-Only	Average	0.81	0.82	0.81	81.72

audio and video modalities. It can be seen that the text modality achieves better accuracy compared to audio and video modalities.

5.1. Discussion

In the comparative experiments described above, all possible multimodal fusion combination including A + V, A + T, T + V, A + V + T were contextually considered. Fig. 10 presents the performance accuracy results of early and late fusion approaches. The results show that early fusion outperforms the late fusion strategy. Furthermore, the full A + V + T multimodal system with early (feature-level) fusion is seen to achieve the highest accuracy compared to late (decision-level) fusion and other early multimodal fusion combinations.

In addition, the comparative experimental results show that multimodal

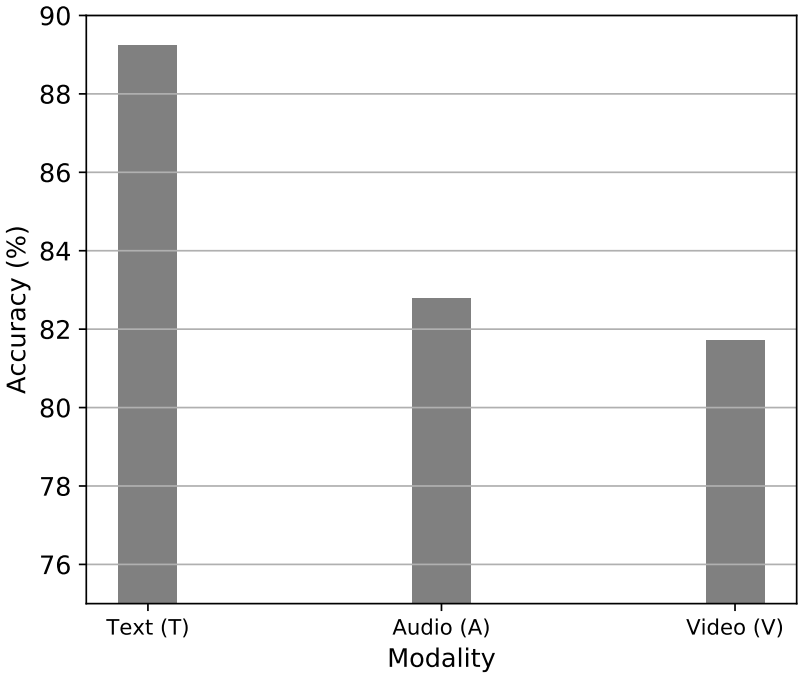

Fig. 9: Comparison of proposed Unimodal Persian Sentiment Analysis

fusion consistently achieves better performance compared to all unimodal sentiment analysis approaches. Amongst the latter, T-only analysis performs better than both other unimodal (V-only, A-only) approaches. The superior performance of the multi-modal (A+V+T) approach compared to the best unimodal (T-only) approach is attributed to the contextual fusion of acoustic, visual and textual features which are able to overcome the limitation of ambiguous words/phrases usage in Persian T-only unimodal sentiment analysis.

Finally, the main limitations of our proposed Persian multimodal sentiment analysis framework are identified below, which need addressed in future research:

- The Persian multimodal sentiment analysis framework cannot detect sub-

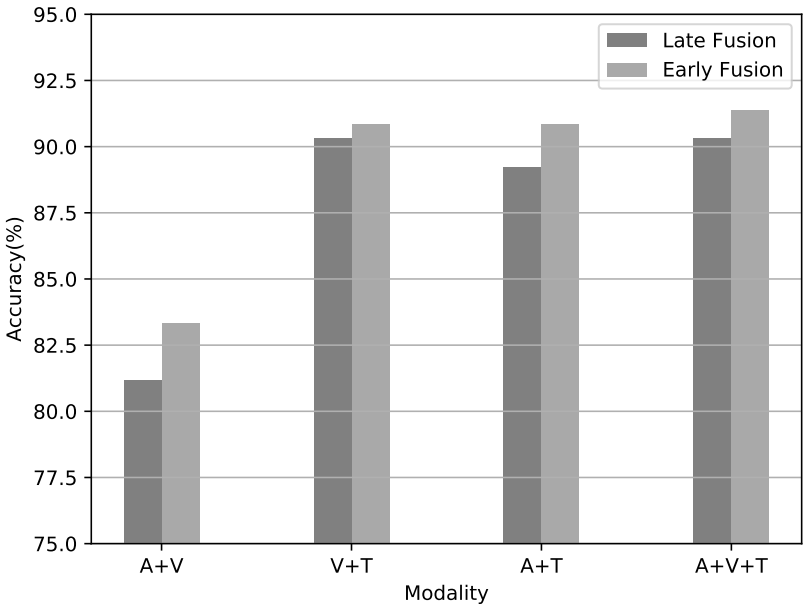

Fig. 10: Early (feature-level) vs Late (decision-level) Multimodal Fusion approaches: Comparison of prediction accuracy, precision, recall and F1-score

jective/objective utterances. For example, بريم قسمتهاي از اين فيلم بّينيم و برگرديم, "Let's go and see some part of the movie and come back". The proposed approach is unable to detect objective utterances.

- The speakers in the videos are not speaking formally i.e. most of the videos comprise informal words to express the speaker's opinion which is difficult for the proposed approach to detect. For example, ياسر بختياري با موزيك رپش حرف دل جوانها رو ميزنه و خيلي تاثير گذر هست "Yasr

Bakhtiari with his rap music is saying what young people want and his music is really effective".

- The current model is restricted to a single speaker and does not work in multi-speaker scenarios. For example, when the video consists of two speakers it cannot detect the polarity of sentences they are simultaneously expressing.

- The number of neutral discourses in the data is very less as compared to positive and negative cases. Therefore, we cannot train our proposed models on neutral comments as deep learning requires a large training dataset. We intend to extend the presented multimodal dataset to include more objective discourse in order to facilitate multimodal subjectivity classification.

It is worth mentioning that the dataset was not translated into English before processing as the translation often fails to work on languages comprising lots of sarcastic and informal words [64]. Therefore, the multimodal approach was directly trained on Persian language without translating the input data into English.

6. Conclusion and Future work

In text-based sentiment analysis, the usual source of information consists of n-gram, word-order, dependency relations and part-of-speech features that often prove inadequate to identify the overall polarity of the natural language sentence. On the other hand, videos comprise multiple modalities including text, audio and visual features which can be contextually exploited to enhance the polarity detection. In this paper, we presented a first of its kind multimodal dataset for Persian language, consisting of utterances and their sentiment polarity extracted from YouTube videos. In addition, a novel Persian multimodal sentiment analysis framework for contextually combining audio, visual and textual features was proposed. Our experimental results demonstrated that the

fusion of Persian audio, visual and textual features outperform all other unimodal classifiers including text-based, audio-based and video-based sentiment analysis models.

In future, we plan to address limitations with current unimodal and multimodal features through contextual extraction of more objective utterances using Persian dependency-rule based approaches [28]. Furthermore, we will explore multilingual approaches to detect polarity in multilingual videos, by exploiting our newly developed language-independent multimodal models [40, 65, 66, 67] and multi-task learning approaches [68].

References

[1] E. Cambria, H. Wang, B. White, Guest editorial: Big social data analysis, Knowledge-Based Systems 69 (2014) 1–2.

[2] M. Grassi, E. Cambria, A. Hussain, F. Piazza, Sentic web: A new paradigm for managing social media affective information, Cognitive Computation 3 (3) (2011) 480–489.

[3] G. Rabby, S. Azad, M. Mahmud, K. Z. Zamli, M. M. Rahman, Teket: a tree-based unsupervised keyphrase extraction technique, Cognitive Computation (2020) 1–23.

[4] E. Cambria, D. Hazarika, S. Poria, A. Hussain, R. Subramanyam, Benchmarking multimodal sentiment analysis, in: International Conference on Computational Linguistics and Intelligent Text Processing, Springer, 2017, pp. 166–179.

[5] X. Zhong, E. Cambria, A. Hussain, Extracting time expressions and named entities with constituent-based tagging schemes, Cognitive Computation 12 (4) (2020) 844–862.

[6] R. Satapathy, E. Cambria, A. Nanetti, A. Hussain, A review of shorthand systems: From brachygraphy to microtext and beyond, Cognitive Computation 12 (4) (2020) 778–792.

[7] C. Angulo, Z. Falomir, D. Anguita, N. Agell, E. Cambria, Bridging cognitive models and recommender systems, Cognitive Computation 12 (2) (2020) 426–427.

[8] E. Ragusa, P. Gastaldo, R. Zunino, E. Cambria, Learning with similarity functions: a tensor-based framework, Cognitive Computation 11 (1) (2019) 31–49.

[9] K. Dashtipour, M. Gogate, A. Adeel, C. Ieracitano, H. Larijani, A. Hussain, Exploiting deep learning for persian sentiment analysis, in: International Conference on Brain Inspired Cognitive Systems, Springer, 2018, pp. 597–604.

[10] K. Dashtipour, M. Gogate, A. Adeel, A. Algarafi, N. Howard, A. Hussain, Persian named entity recognition, in: 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2017, pp. 79–83.

[11] K. Dashtipour, M. Gogate, A. Adeel, A. Hussain, A. Alqarafi, T. Durrani, A comparative study of persian sentiment analysis based on different feature combinations, in: International Conference in Communications, Signal Processing, and Systems, Springer, 2017, pp. 2288–2294.

[12] A. Hussain, A. Tahir, Z. Hussain, Z. Sheikh, M. Gogate, K. Dashtipour, A. Ali, A. Sheikh, Artificial intelligence-enabled analysis of uk and us public attitudes on facebook and twitter towards covid-19 vaccinations, medRxiv.

[13] W. Jones, A grammar of the Persian language, Vol. 5, John Stockdale, 1807.

[14] S. L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: from formal to informal and scarce resource languages, Artificial Intelligence Review 48 (4) (2017) 499–527.

[15] R. Ahmed, K. Dashtipour, M. Gogate, A. Raza, R. Zhang, K. Huang, A. Hawalah, A. Adeel, A. Hussain, Offline arabic handwriting recognition

using deep machine learning: A review of recent advances, in: International Conference on Brain Inspired Cognitive Systems, Springer, 2019, pp. 457–468.

[16] S. K. Yadav, M. Bhushan, S. Gupta, Multimodal sentiment analysis: Sentiment analysis using audiovisual format, in: 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2015, pp. 1415–1419.

[17] F. Jiang, B. Kong, J. Li, K. Dashtipour, M. Gogate, Robust visual saliency optimization based on bidirectional markov chains, Cognitive Computation.

[18] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognition Letters 125 (264-270).

[19] E. Cambria, N. Howard, J. Hsu, A. Hussain, 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, 2013, pp. 108–117.

[20] H.-N. Tran, E. Cambria, Ensemble application of ELM and GPU for real-time multimodal sentiment analysis, Memetic Computing 10 (2018) 3–13.

[21] E. Cambria, A. Hussain, Sentic album: Content-, concept-, and context-based online personal photo management system, Cognitive Computation 4 (4) (2012) 477–496.

[22] S. Poria, H. Peng, A. Hussain, N. Howard, E. Cambria, Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis, Neurocomputing 261 (2017) 217–230.

[23] E. Cambria, Y. Li, F. Z. Xing, S. Poria, K. Kwok, Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis, in:

Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 105–114.

[24] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.

[25] K. Dashtipour, A. Hussain, Q. Zhou, A. Gelbukh, A. Y. Hawalah, E. Cambria, Persent: A freely available persian sentiment lexicon, in: International Conference on Brain Inspired Cognitive Systems, Springer, 2016, pp. 310–320.

[26] K. Dashtipour, A. Raza, A. Gelbukh, R. Zhang, E. Cambria, A. Hussain, Persent 2.0: Persian sentiment lexicon enriched with domain-specific words, in: International Conference on Brain Inspired Cognitive Systems, Springer, 2019, pp. 497–509.

[27] K. Dashtipour, C. Ieracitano, F. C. Morabito, A. Raza, A. Hussain, An ensemble based classification approach for persian sentiment analysis, in: Progresses in Artificial Intelligence and Neural Systems, Springer, 2020, pp. 207–215.

[28] K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, A. Hussain, A hybrid persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks, Neurocomputing 380 (2020) 1–10.

[29] A. Tahir, G. Morison, D. A. Skelton, R. M. Gibson, A novel functional link network stacking ensemble with fractal features for multichannel fall detection, Cognitive Computation 12 (5) (2020) 1024–1042.

[30] X. Jiang, T. Yan, J. Zhu, B. He, W. Li, H. Du, S. Sun, Densely connected deep extreme learning machine algorithm, Cognitive Computation 12 (5) (2020) 979–990.

[31] B. Elayeb, A. Chouigui, M. Bounhas, O. B. Khiroun, Automatic arabic text summarization using analogical proportions, Cognitive Computation 12 (5) (2020) 1043–1069.

[32] E. Cambria, T. Mazzocco, A. Hussain, Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining, Biologically Inspired Cognitive Architectures 4 (2013) 41–53.

[33] A. Adeel, M. Gogate, S. Farooq, C. Ieracitano, K. Dashtipour, H. Larijani, A. Hussain, A survey on the role of wireless sensor networks and iot in disaster management, in: Geological disaster monitoring based on sensor networks, Springer, 2019, pp. 57–66.

[34] F. Jiang, B. Kong, J. Li, K. Dashtipour, M. Gogate, Robust visual saliency optimization based on bidirectional markov chains, Cognitive Computation (2020) 1–12.

[35] S. M. Asad, K. Dashtipour, S. Hussain, Q. H. Abbasi, M. A. Imran, Travelers-tracing and mobility profiling using machine learning in railway systems, in: 2020 International Conference on UK-China Emerging Technologies (UCET), IEEE, 2020, pp. 1–4.

[36] Z. Yu, P. Machado, A. Zahid, A. M. Abdulghani, K. Dashtipour, H. Heidari, M. A. Imran, Q. H. Abbasi, Energy and performance trade-off optimization in heterogeneous computing via reinforcement learning, Electronics 9 (11) (2020) 1812.

[37] M. Gogate, A. Adeel, A. Hussain, Deep learning driven multimodal fusion for automated deception detection, in: Computational Intelligence (SSCI), 2017 IEEE Symposium Series on, IEEE, 2017, pp. 1–6.

[38] C. Ieracitano, N. Mammone, A. Hussain, F. C. Morabito, A novel multimodal machine learning based approach for automatic classification of eeg recordings in dementia, Neural Networks 123 (2020) 176–190.

[39] D. Hu, F. Nie, X. Li, Deep multimodal clustering for unsupervised audiovisual learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9248–9257.

[40] M. Gogate, K. Dashtipour, A. Adeel, A. Hussain, Cochleanet: A robust language-independent audio-visual model for speech enhancement, Information Fusion.

[41] A. Adeel, M. Gogate, A. Hussain, W. M. Whitmer, Lip-reading driven deep learning approach for speech enhancement, IEEE Transactions on Emerging Topics in Computational Intelligence.

[42] N. Howard, A. Adeel, M. Gogate, A. Hussain, Deep cognitive neural network (dcnn), uS Patent App. 16/194,721 (May 23 2019).

[43] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web, in: Proceedings of the 13th international conference on multimodal interfaces, ACM, 2011, pp. 169–176.

[44] S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2539–2544.

[45] V. P. Rosas, R. Mihalcea, L.-P. Morency, Multimodal sentiment analysis of spanish online videos, IEEE Intelligent Systems 28 (3) (2013) 38–45.

[46] A. S. Alqarafi, A. Adeel, M. Gogate, K. Dashitpour, A. Hussain, T. Durrani, Toward's arabic multi-modal sentiment analysis, in: International Conference in Communications, Signal Processing, and Systems, Springer, 2017, pp. 2378–2386.

[47] M. B. Dastgheib, S. Koleini, F. Rasti, The application of deep learning in persian documents sentiment analysis, International Journal of Information Science and Management (IJISM) 18 (1) (2020) 1–15.

[48] M. Farahani, M. Gharachorloo, M. Farahani, M. Manthouri, Parsbert: Transformer-based model for persian language understanding, arXiv preprint arXiv:2005.12515.

[49] Y. Wang, M. Huang, L. Zhao, et al., Attention-based lstm for aspect-level sentiment classification, in: Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 606–615.

[50] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext. zip: Compressing text classification models, arXiv preprint arXiv:1612.03651.

[51] Y. Kim, Convolutional neural networks for sentence classification, In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar, Association for Computational Linguistics.

[52] R. Dehkharghani, B. Yanikoglu, D. Tapucu, Y. Saygin, Adaptation and use of subjectivity lexicons for domain dependent sentiment classification, in: 2012 IEEE 12th International Conference on Data Mining Workshops, IEEE, 2012, pp. 669–673.

[53] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, 2017, pp. 427–431.

[54] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, M. Pantic, A survey of multimodal sentiment analysis, Image and Vision Computing 65 (2017) 3–14.

[55] Z. Wang, S. Ho, E. Cambria, A review of emotion sensing: Categorization models and algorithms, Multimedia Tools and Applications 79 (2020) 35553–35582.

[56] Y. Susanto, A. Livingstone, B. C. Ng, E. Cambria, The hourglass model revisited, IEEE Intelligent Systems 35 (5) (2020) 96–102.

[57] M. A. Turk, A. P. Pentland, Face recognition using eigenfaces, in: Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1991, pp. 586–591.

[58] P. Werner, F. Saxen, A. Al-Hamadi, Landmark based head pose estimation benchmark and method, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 3909–3913.

[59] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, K. M. Prkachin, Automatically detecting pain in video through facial action units, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41 (3) (2010) 664–674.

[60] S. K. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Computing Surveys (CSUR) 47 (3) (2015) 43.

[61] C. G. Snoek, M. Worring, A. W. Smeulders, Early versus late fusion in semantic video analysis, in: Proceedings of the 13th annual ACM international conference on Multimedia, ACM, 2005, pp. 399–402.

[62] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, P. Natarajan, Multimodal feature fusion for robust event detection in web videos, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1298–1305.

[63] M. Zuckerman, J. A. Hall, R. S. DeFrank, R. Rosenthal, Encoding and decoding of spontaneous and posed facial expressions., Journal of Personality and Social Psychology 34 (5) (1976) 966.

[64] G. Lazard, et al., A grammar of contemporary Persian, Mazda Publishers Costa Mesa, CA, 1992.

[65] M. Gogate, K. Dashtipour, P. Bell, A. Hussain, Deep neural network driven binaural audio visual speech separation, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[66] M. Gogate, K. Dashtipour, A. Hussain, Visual speech in real noisy environments (vision): A novel benchmark dataset and deep learning-based baseline system, Proc. Interspeech 2020 (2020) 4521–4525.

[67] M. Gogate, A. Adeel, K. Dashtipour, P. Derleth, A. Hussain, Av speech enhancement challenge using a real noisy corpus, arXiv preprint arXiv:1910.00424.

[68] F. Xiong, B. Sun, X. Yang, H. Qiao, K. Huang, A. Hussain, Z. Liu, Guided policy search for sequential multitask learning, IEEE Transactions on Systems, Man, and Cybernetics: Systems 49 (1) (2018) 216–226.