Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification

Mahesh G. Huddar*

Department of Computer Science and Engineering,
Hirasugar Institute of Technology,
Nidasoshi, Belagavi, 591236, India
Email: mailtomgh1@gmail.com
*Corresponding author

Sanjeev S. Sannakki and Vijay S. Rajpurohit

Department of Computer Science and Engineering,
Gogte Institute of Technology,
Belagavi, 590008, India
Email: sannakkisanjeev@gmail.com
Email: vijaysr2k@yahoo.com

Abstract: Multimodal affective computing has become a popular research area, due to the availability of a large amount of multimodal content. Feature alignment between the modalities and multimodal fusion are the most important issues in multimodal affective computing. To address these issues, the proposed model extracts the features at word-level and forced alignment is used to understand the time-dependent interaction among the modalities. The contextual information among the words of an utterance and between the nearby utterances is extracted using bidirectional long short term memory (LSTM). Weighted pooling based attention model is used to select the important features within the modalities and importance of each modality. Information from multiple modalities is fused using a cross-modality fusion technique. The performance of the proposed model was tested on two standard datasets such as IEMOCAP and CMU-MOSI. By incorporating the word-level features, feature alignment, and cross-modality fusion, the proposed architecture outperforms the baselines in terms of classification accuracy.

Keywords: affective computing; attention model; contextual fusion; cross-modality fusion; feature alignment; computer vision; deep learning; bidirectional recurrent neural network; sentiment analysis.

Reference to this paper should be made as follows: Huddar, M.G., Sannakki, S.S. and Rajpurohit, V.S. (2020) ‘Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification’, *Int. J. Intelligent Engineering Informatics*, Vol. 8, No. 1, pp.1–18.

Biographical notes: Mahesh G. Huddar is an Assistant Professor in the Department of Computer Science and Engineering at Hirasugar Institute of Technology, Nidasoshi, Belagavi, India and he is currently pursuing PhD in Visvesvaraya Technological University, Belagavi, India. He received his Master of Technology and Bachelor of Engineering degrees from the

Visvesvaraya Technological University, Belagavi, India in 2014 and 2008, respectively. He has published a good number of papers in journals, international, and national conferences. His main research interests include machine learning, deep learning, multimodal sentiment analysis, and emotion detection.

Sanjeev S. Sannakki has completed his PhD in Image Processing and Data Mining from VTU Belagavi. His career spans over two decades in the field of teaching, research and other diversified academics. He is currently working as a Professor in the Department of Computer Science and Engineering, Gogte Institute of Technology, Belgaum. Currently, he is shouldering the responsibility of Head of the Research Center. He has published several papers in reputed national/international conferences and journals. He is also guiding the research scholars and UG/PG students of VTU.

Vijay S. Rajpurohit is working as a Professor in the Department of Computer Science and Engineering at Gogte Institute of Technology, Belgaum, Karnataka, India. He pursued his BE in Computer Science and Engineering from Karnataka University Dharwad, MTech from N.I.T.K Surathkal, and a PhD from Manipal University, Manipal in 2009. His research areas include image processing, cloud computing, and data analytics. He has published a good number of papers in journals, International, and National conferences. He is the reviewer for a few international journals and conferences. He is the associate editor for two international journals and a Senior Member of International Association of CS and IT.

1 Introduction

Due to the rapid advancements in world-wide-web (WWW), increased usage of Smartphone's and affordable internet availability led the user to post their reviews about a product or an entity in an audio-visual format. People share their opinion about a newly released movie or product or service or tourist place or any new entity in an audio-visual format on Twitter, YouTube, Flickr, Facebook (Poria et al., 2017a). This enables business industries to analyse the sentiment or complaint or suggestions of customers to increase the profit by improving the quality of service or product etc. The main advantage of analysing audio-visual (textual, audio and visual modalities) data over only textual modality for affective computing is the availability of acoustic and body gestures in video content. Also, few modalities have different importance in multimodal affective computing, for example, audio and video modality play a more important role in emotion classification while less in sentiment analysis, also textual modality has more importance in sentiment analysis. The combination of different audio-visual-textual modalities helps us to build more accurate and robust multimodal affective computing models (Huddar et al., 2019a). Basically, sentiment analysis and emotion classification models extract features from the audio-visual modalities, use either feature (early) level fusion (Pérez-Rosas et al., 2013) or model-based fusion (Du et al., 2018) or decision (late) level fusion (Huddar et al., 2018) to merge the information extracted from audio-visual modalities.

In multimodal affective computing, video is divided into segments (utterances) of length 5–15 seconds and an average of 12 words per utterance. The features are extracted at the utterance level from multiple modalities and merged to get the final multimodal feature vector. The utterance level analysis posts the promising results in multimodal affective computing (Poria et al., 2017b). The problem with utterance level sentiment analysis and emotion classification is the alignment or synchronisation of features across the modalities cannot be accommodated. Also, the emotion or sentiment of an utterance (segment) may depend on the nearby utterances. These problems can be addressed by extracting contextual word-level features and using cross-modality fusion among the multiple modalities.

The major contributions of this proposed work are:

- a word-level multimodal contextual feature extraction technique that is able to extract and align contextual features across modalities at word-level
- a hierarchical cross-modality fusion that can identify modality-specific important features from individual modality and the importance of each modality before fusion
- the results demonstrate that the proposed attention-based word-level feature extraction and cross-modality fusion method outperform the state-of-the-art methods in terms of classification accuracy.

The rest of the study contains, Section 2 briefly discusses the recent literature in multimodal affective computing. The proposed attention-based word-level feature extraction and cross-modality fusion are discussed in Section 3. In Section 4, the experimental results of the proposed method are discussed. Finally, the study is concluded with a discussion on future work in Section 5.

2 Related work

Usually, sentiment and emotion classification is performed on textual data (Abanda, 2017). The business world is using sentiment analysis and emotion classification to address many issues such as improving the quality of products and services (Mars and Gouider, 2017), to forecast upcoming political results (Ramteke et al., 2016), growth in the financial sector (Li et al., 2014), improve quality of e-tourism (Kirilenko et al., 2018) and e-health (Gohil et al., 2018), to predict movie box office collection (Nagamma et al., 2015), etc. The availability of a large amount of multimodal data on the internet led researchers to think beyond textual sentiment analysis and emotion classification (Huddar et al., 2019b). In the early literature low-level handcrafted (manually extracted) features such as lexicon representation for textual data, low-level descriptors for speech and facial expressions for video modalities (Rosas et al., 2013) or ontologies (Mahmoud et al., 2018) or sentiment lexicons (Mohammad et al., 2013) were considered. Instead of simplistic fusion techniques such as feature level fusion (Pérez-Rosas et al., 2013) or model-based fusion (Du et al., 2018) or decision level fusion (Huddar et al., 2018), shallow fusion techniques were proposed to understand the complex correlation between modalities.

In recent literature Bayesian filtering (Savran et al., 2012), non-trainable tensor-based (Zadeh et al., 2017) and memory-based fusion networks (Zadeh et al., 2018), and ensemble SVM trees (Rozgić et al., 2012) were proposed. As the manual handcrafted features cannot extract complex hidden correlations between the modalities, deep learning-based models were proposed to extract high-level features. Recurrent neural networks such as long short term memory (LSTM) and gated recurrent unit (GRU) were used to extract visual and speech-related features (Gu et al., 2018) and convolutional neural network-based models were used to extract visual features. Deep learning-based feature extraction models automatically learn patterns among the data and extract features that are general and more accurate compared to handcrafted manually extracted features. Bidirectional LSTMs and GRUs were also used to extract long-term dependencies and contextual information among the utterances. Researchers combine high-level features at feature (early) level fusion or decision level fusion such as voting or averaging to get the combined representations of modalities (Wöllmer et al., 2013).

Most recently researchers focused on building model-level fusion techniques, for generating a common representation of feature vector for different modalities (Trabelsi et al., 2017). These works fuse raw features using gated embedded fusion technique by ignoring the temporal information. To address these problems the proposed method extracts word-level features of individual modalities. Later, bidirectional LSTM is used to understand the relatedness among the word-level features among the modalities. Finally, important word-level features (modality-specific features) are selected using a weighted pooling based attention model before fusion. In multimodal affective computing different modalities like text, video and audio may have different importance. For example, textual features play an important role in sentiment analysis, but less in emotion recognition, similarly acoustic features more in emotion recognition but less in sentiment analysis. Hence in the proposed method weighted pooling based attention method is used to select the importance of each modality (cross-modality features) before fusion. Finally, modality-specific features and cross-modality features are fused to get bimodal and subsequently trimodal representation of feature vectors. The proposed architecture is demonstrated on two publically published datasets (see Section 3.1) and show that the proposed word-level feature extraction and cross-modality fusion using attention-based model achieves better results in terms of classification accuracy (see Section 4).

3 Proposed methodology

In this section, the proposed attention-based word-level contextual feature extraction, alignment across modalities and cross-modality fusion are discussed in detail. The brief overview of the proposed technique is:

- first, word-level audio, textual and visual features are extracted, next important contextual features are selected using bidirectional LSTM and weighted pooling based attention model (see Section 3.2)
- *input layer*: unimodal word-level contextual feature vectors
- *bimodal fusion*: obtain bimodal contextual feature vector by fusing two-two modalities at a time by considering important features from individual modalities and the importance of cross modalities (see Section 3.4.1)
- *trimodal fusion*: obtain trimodal contextual feature vector by fusing all modalities by considering important features from individual modalities and the importance of cross modalities (see Section 3.4.2)

- *biLSTM with attention layer*: bimodal and the trimodal contextual feature vector is obtained using bidirectional LSTM and important features using weighted pooling based attention model
- *output layer*: the trimodal contextual feature vector is fed to a softmax output layer for affective computing (see Section 3.5).

3.1 Dataset used

The IEMOCAP (Busso et al., 2008) and CMU-MOSI (Zadeh et al., 2016) are two publicly available standard multimodal datasets, used for multimodal emotion recognition and multimodal sentiment classification respectively.

3.1.1 IEMOCAP

IEMOCAP database is a dyadic acted multi-speaker and a multimodal dataset containing nearly 12 hours of text, audio and video conversations among male and female participants. The conversations are divided into small segments/utterances each of duration 3–15 seconds. Multiple assessors evaluate the utterances manually and assign an affective label. The voted result from multiple assessors is used as a label for an utterance. To be consistent with the recent literature, out of ten available affective labels four labels such as angry, sad, happy (excitement) and neutral were considered for the study. The train/test split of the IEMOCAP dataset is shown in Table 1.

Table 1 Train/test split of IEMOCAP dataset

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

3.1.2 CMU-MOSI

CMU-MOSI database is a multi-speaker and multimodal dataset containing 93 English review videos. The dataset is divided into 2199 small segments/utterances, with an average length of 12 words and 4.2 seconds duration per utterances. The five assessors manually labelled each of the utterances with a score between –3 (strongly negative) and +3 (strongly positive). To be in line with the recent literature, voted result from multiple assessors is used as the label and only positive and negative labelled utterances were considered for the study. The train/test split of the CMU-MOSI dataset is shown in Table 2.

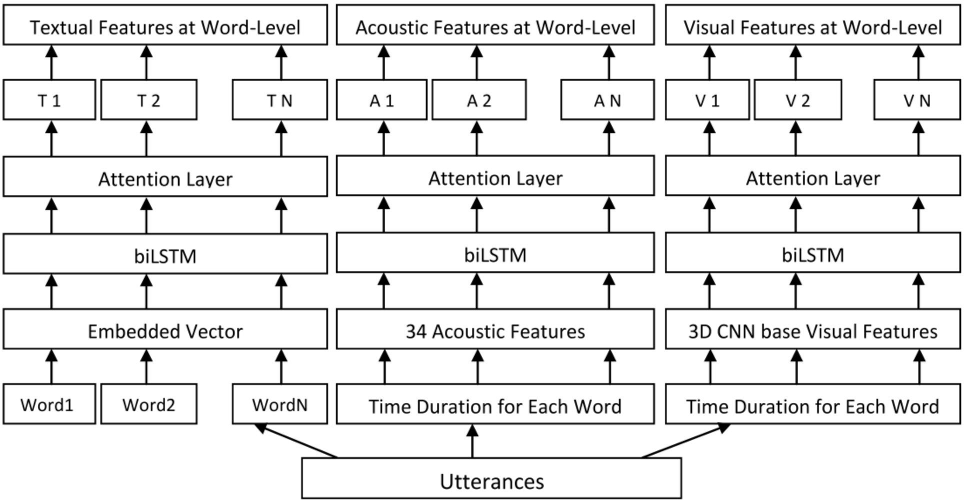
Table 2 Train/test split of CMU-MOSI dataset

	Positive	*Negative*
Train	709	738
Test	467	285

3.2 Feature extraction

Word-level feature extraction and feature alignment among the modalities are the fundamental requirements of the proposed approach. In this approach, features are extracted at word-level and synchronised across modalities. Word-level feature extraction is used as words are the basic unit in human conversation. Merging the features at word level across modalities becomes difficult, as different modalities have different sampling rates. To address this issue there are two options, first, down sample the data from different modalities to common rate, second, extract the features from individual modalities at their normal sample rate and use the encoder to create a common feature vector. As down sampling leads to loss of information, the second option is considered for further experiments. The overview of word-level contextual feature extraction and alignment strategy is shown in Figure 1 and discussed in detail in further sections.

Figure 1 Overview of word-level contextual feature extraction and alignment


Textual Features at Word-Level
Acoustic Features at Word-Level
Visual Features at Word-Level
T 1
T 2
T N
A 1
A 2
A N
V 1
V 2
V N
Attention Layer
Attention Layer
Attention Layer
biLSTM
biLSTM
biLSTM
Embedded Vector
34 Acoustic Features
3D CNN base Visual Features
Word1
Word2
WordN
Time Duration for Each Word
Time Duration for Each Word
Utterances

3.2.1 Textual feature extraction

The three-step approach is followed to extract textual word-level features. Firstly, each word of an utterance is embedded using pre-trained Word2Vec dictionary (Mikolov et al., 2013), to prepare word-level representation for textual data. Let N is the number of utterances, M is the number of words in a given utterance, w_{ij} is the jth word in ith utterance and E is the Word2Vec embedding matrix, then the embedded vector T_{ij} of jth word in ith utterance is represented by,

$$T_{ij} = Ew_{ij}, \quad i \in [1, N] \text{ and } j \in [1, M] \tag{1}$$

Secondly, as CNN based feature extraction uses a fixed size window, which restricts feature extraction from varying sized utterances (Poria et al., 2015). Also, contextual information among the words of an utterance cannot be extracted using CNN models. To overcome these problems, bidirectional LSTM is used to extract the contextual information at word-level.

$$\overrightarrow{t_{ij}}\overleftarrow{t_{ij}} = biLSTM(T_{ij}), \quad i \in [1, N] \text{ and } j \in [1, M] \tag{2}$$

where biLSTM is the bidirectional LSTM, $\overrightarrow{t_{ij}}$ and $\overleftarrow{t_{ij}}$ are the forward and backward contextual information states of input text respectively. The final word-level feature vector t_{ij} is obtained by combining $\overrightarrow{t_{ij}}$ and $\overleftarrow{t_{ij}}$.

Thirdly, the attention-based strategy is used to extract informative features at word-level. Weighted pooling based attention layer generates one-dimensional weight vector over the word-level contextual information.

As described in Tao and Liu (2018), the textual energy te_{ij} and weight distribution td_{ij} of the word-level feature vector t_{ij} are computed as,

$$te_{ij} = \tanh\left(W_t t_{ij} + b\right), \quad i \in [1, N] \text{ and } j \in [1, M] \tag{3}$$

$$td_{ij} = \frac{\exp\left(te_{ij}\right)}{\sum_{k=1}^{M} \exp\left(te_{ik}\right)} \tag{4}$$

where W_t is the learnable parameter.

The word-level feature vector t_{ij} and text weight distribution td_{ij} are used as input in understanding word-level fusion across modalities in further steps.

3.2.2 Audio feature extraction

The three-step approach is followed to extract word-level acoustic feature extraction. Firstly, the time interval (frame) for each word in the audio file is calculated using the Sakoe-Chiba band dynamic time warping (DTW) (Sakoe and Chiba, 1978). All the sentences are zero-padded to make all sentences of equal length (*L* frames).

$$w_{ik} \leftrightarrow A_{ij}, \quad i \in [1, N], k \in [1, M] \text{ and } j \in [1, L] \tag{5}$$

where w_{ik} is the *k*th word in *i*th utterance, A_{ij} is the *j*th acoustic time interval equivalent to *k*th word in *i*th utterance. Secondly, as stated in Huddar et al. (2019b) 34 audio features such as "13 energy-based Mel-frequency cepstral coefficients (MFCC) features, eight time-spectral features like short-term entropy, spectral spread, zero-crossing rate, spectral roll-off, spectral centroid, short-term energy, spectral entropy, and flux and 13 chroma-based features" are extracted. Bidirectional LSTM is used to extract contextual information among word-level acoustic features.

$$\overrightarrow{a_{ij}}\overleftarrow{a_{ij}} = biLSTM(A_{ij}), \quad i \in [1, N] \text{ and } j \in [1, L] \tag{6}$$

where $\overrightarrow{a_{ij}}$ and $\overleftarrow{a_{ij}}$ are the forward and backward contextual information states of word-level acoustic input respectively. The final word-level acoustic feature vector a_{ij} is obtained by combining $\overrightarrow{a_{ij}}$ and $\overleftarrow{a_{ij}}$. Thirdly, similar to text modality, the attention-based strategy is used to extract informative word-level acoustic features. Weighted pooling based attention layer generates a one-dimensional weight vector over the acoustic word-level contextual information. As described in Tao and Liu (2018), the word-level acoustic energy ae_{ij} and weight distribution ad_{ij} of the word-level acoustic feature vector a_{ij} are computed as,

$$ae_{ij} = \tanh\left(W_t a_{ij} + b\right), i \in [1, N] \text{ and } j \in [1, L] \tag{7}$$

$$ad_{ij} = \frac{\exp\left(ae_{ij}\right)}{\sum_{k=1}^{L} \exp\left(ae_{ik}\right)} \tag{8}$$

where W_t is the learnable parameter.

The acoustic word-level feature vector a_{ij} and audio weight distribution ad_{ij} are used as input in understanding acoustic word-level fusion across modalities in further steps.

3.2.3 Visual feature extraction

Similar to text and audio modalities, the three-step approach is used to extract visual features. Firstly, similar to acoustic word-level feature extraction, the time duration for each word in video content is calculated. All the sentences are zero-padded to make all sentences of equal length (L frames).

$$w_{ik} \leftrightarrow V_{ij}, \quad i \in [1, N], j \in [1, M] \text{ and } j \in [1, L] \tag{9}$$

where w_{ik} is the kth word in ith utterance, V_{ij} is the jth visual time interval equivalent to kth word in ith utterance. Secondly, visual features are extracted using 3D-CNN networks. 3D-CNN networks are used and proved their importance in object detection, recognition, and classification (Abdellaoui and Douik, 2018). Using 3D-CNN networks frame-level visual features were extracted in Poria et al. (2017b). The study replicates the process of extracting visual features at the frame level. Bidirectional LSTM is used to extract contextual information among word-level visual features.

$$\overrightarrow{v_{ij}}\overleftarrow{v_{ij}} = biLSTM(V_{ij}), \quad i \in [1, N] \text{ and } j \in [1, L] \tag{10}$$

where $\overrightarrow{v_{ij}}$ and $\overleftarrow{v_{ij}}$ are the forward and backward contextual information states of word-level visual input respectively. The final word-level visual feature vector v_{ij} is obtained by combining $\overrightarrow{v_{ij}}$ and $\overleftarrow{v_{ij}}$. Thirdly, similar to text and audio modality, the attention-based strategy is used to extract informative word-level visual features. Weighted pooling based attention layer generates a one-dimensional weight vector over the visual word-level contextual information. As described in Tao and Liu (2018), the word-level visual energy ve_{ij} and weight distribution vd_{ij} of the word-level visual feature vector v_{ij} are computed as,

$$ve_{ij} = \tanh\left(W_t v_{ij} + b\right), \quad i \in [1, N] \text{ and } j \in [1, L] \tag{11}$$

$$vd_{ij} = \frac{\exp\left(ve_{ij}\right)}{\sum_{k=1}^{L} \exp\left(ve_{ik}\right)} \tag{12}$$

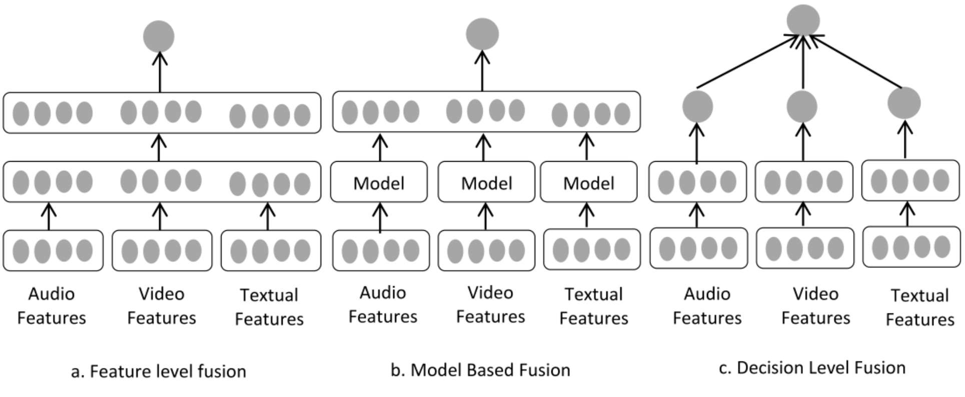
where W_t is the learnable parameter.

The acoustic word-level feature vector v_{ij} and visual weight distribution vd_{ij} are used as input in understanding visual word-level fusion across modalities in further steps.

3.3 Problem with early, model-based and late fusion

Recent literature in multimodal affective computing uses either early fusion (also known as feature concatenation) (Pérez-Rosas et al., 2013) or model-based fusion (Du et al., 2018) or decision fusion (also known as late fusion) (Huddar et al., 2018). In early or feature level fusion, features from different modalities are concatenated. Then the classification model is built using the concatenated feature vector. In model-based fusion, classification models are built using feature vectors from individual modalities, then the models are combined using concatenation, finally, the result from the combined model is used to assign a class label. In late or decision level fusion, the classification model is built using feature vectors from individual modalities then the results from the classification models are combined to get the final result. Figure 2 shows the early, model-based and late fusion techniques. The feature (early) level fusion considers each modality contributes equally in the fusion. Late (decision) level fusion uses an ensemble approach such as voting or averaging to make the final prediction. Also, these simplistic fusion techniques cannot extract redundant and contextual information between the utterances of the multimodal data. To address these issues, attention-based word-level contextual feature extraction, feature alignment, and cross-modality fusion approach is proposed for multimodal affective computing using bidirectional LSTM. Firstly, word-level textual, audio and visual contextual features are extracted. Word-level features are concatenated to form utterance level features. Next modality-specific attention is used to select modality-specific important features. Modality score is calculated to understand the importance of each modality. Finally, Cross modality fusion is used to fuse the modality-score and modality-specific features.

Figure 2 Multimodality fusion strategies (a) feature (early) level fusion: feature concatenation from multiple modalities; (b) model-based fusion and (c) decision (late) level fusion: prediction fusion from multiple modalities



3.4 Attention-based cross-modality fusion

3.4.1 Bimodal fusion

The word-level textual, audio and visual feature vectors and weight distributions are the input to bimodal fusion. Bimodal fusion has four steps; firstly, we combine the unimodal word-level bidirectional contextual states (t_{ij}, a_{ij}, and v_{ij}) and weight distributions (td_{ij}, ad_{ij}, and vd_{ij}) to form word-level textual, acoustic and visual representations,

$$T_{ij} = t_{ij} td_{ij}, \quad i \in [1, N] \text{ and } j \in [1, M] \tag{13}$$

$$A_{ij} = a_{ij} ad_{ij}, \quad i \in [1, L] \text{ and } j \in [1, M] \tag{14}$$

$$V_{ij} = v_{ij} vd_{ij}, \quad i \in [1, L] \text{ and} j \in [1, M] \tag{15}$$

where T_{ij}, A_{ij} and V_{ij} are word-level textual, audio and visual representations respectively.

Secondly, combine the unimodal contextual states two at a time (Text + Audio, Text + Video, and Audio + Video) to form the bimodal representations,

$$TA_{ij} = T_{ij} A_{ij}, \quad i \in [1, N] \text{ and } j \in [1, M] \tag{16}$$

$$TV_{ij} = T_{ij} V_{ij}, \quad i \in [1, N] \text{ and } j \in [1, M] \tag{17}$$

$$AV_{ij} = A_{ij} V_{ij}, \quad i \in [1, N] \text{ and } j \in [1, M] \tag{18}$$

Further, we employ biLSTM with modality-specific attention to extract the bimodal contextual information and defined as,

$$ta\_s_{ij} = biLSTM\left(TA_{ij}\right), \quad i \in [1, N] \text{ and } j \in [1, M] \tag{19}$$

$$tv\_s_{ij} = biLSTM\left(TV_{ij}\right), \quad i \in [1, N] \text{ and } j \in [1, M] \tag{20}$$

$$av\_s_{ij} = biLSTM\left(AV_{ij}\right), \quad i \in [1, N] \text{ and } j \in [1, M] \tag{21}$$

Thirdly, we apply cross-modality attention to find the modality score to denote the importance of each modality among the combination of modalities in the fusion.

Fourth, soft-attention is used to combine the results of modality-specific attention and cross-modality attention to get bimodal fused feature vector (say $\mathbf{ta_{ij}}$, $\mathbf{tv_{ij}}$ **and** $\mathbf{av_{ij}}$).

3.4.2 Trimodal fusion

Bimodal feature vectors are hierarchically fused to get the trimodal feature vector. Trimodal fusion has three steps; firstly, we combine the bimodal word-level bidirectional contextual states (ta_{ij}, tv_{ij}, and av_{ij}) to form word-level trimodal representations,

$$TAV_{ij} = \left(ta_{ij}, tv_{ij}, av_{ij}\right), \quad i \in [1, N] \text{ and } j \in [1, M] \tag{22}$$

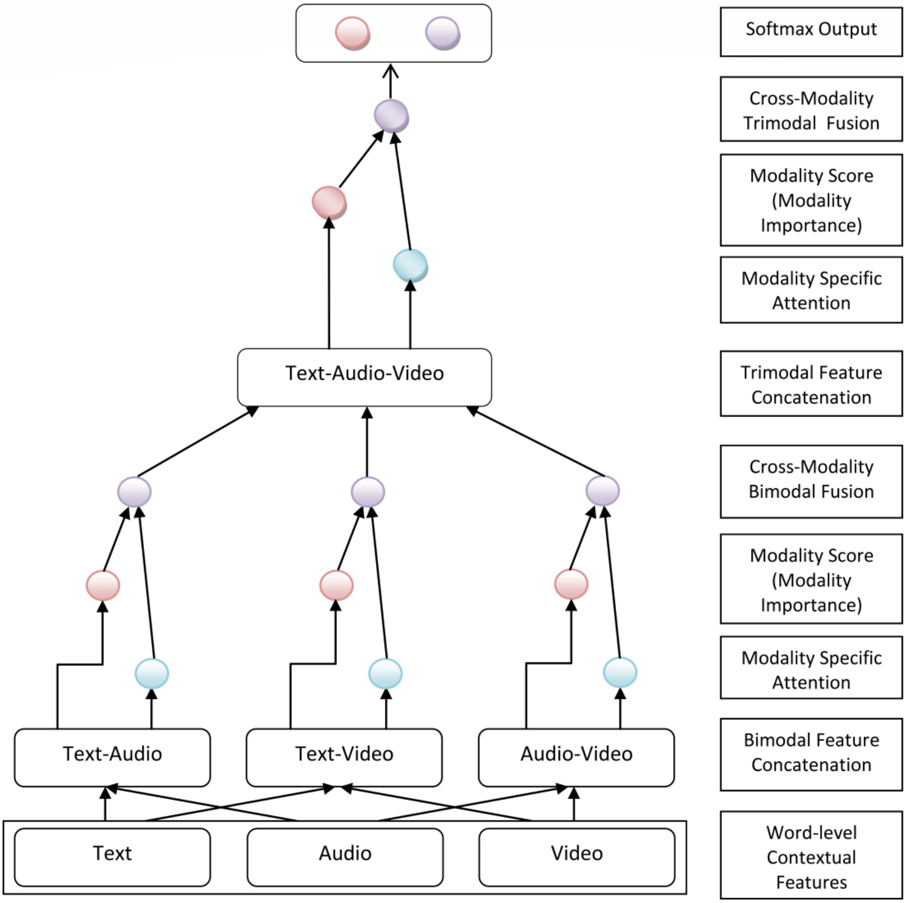
Further, we employ biLSTM with modality-specific attention to extract the trimodal contextual information and defined as,

$$tav\_s_{ij} = biLSTM\left(TAV_{ij}\right), \quad i \in [1, N] \text{ and } j \in [1, M] \tag{23}$$

Secondly, we apply cross-modality attention to find the modality score to denote the importance of each modality among the combination of modalities in the fusion.

Thirdly, soft-attention is used to combine the results of modality-specific attention and cross-modality attention to get trimodal fused feature vector (say tav_{ij}).

Figure 3 Proposed attention-based word-level contextual feature extraction and cross-modality fusion (see online version for colours)


Softmax Output
Cross-Modality Trimodal Fusion
Modality Score (Modality Importance)
Modality Specific Attention
Trimodal Feature Concatenation
Cross-Modality Bimodal Fusion
Modality Score (Modality Importance)
Modality Specific Attention
Bimodal Feature Concatenation
Word-level Contextual Features
Text-Audio-Video
Text-Audio
Text-Video
Audio-Video
Text
Audio
Video

3.5 Training and classification

As the dataset is labelled at the utterance level, the training and classification are performed at the utterance level. Word-level trimodal contextual feature vector tav_{ij} of ith utterance acts as the input to the softmax classifier. The softmax classifier is represented by,

$$p(y \mid U) = \text{softmax}\left(w^{(s)} tav_{ij} + b^{(s)}\right), \quad i \in [1, N] \text{ and } j \in [1, M] \tag{24}$$

where $w^{(s)}$ and $b^{(s)}$ are the weight matrix and bias matrices respectively.

The final class label $\hat{y}$ for the testing utterance, i.e., happy (excitement), angry, neutral and sadness for affective computing and positive and negative for sentiment classification is represented as,

$$\hat{y} = \arg\max_{y} p(y \mid U) \tag{25}$$

where p is a predicted value of the testing utterance.

Table 3 Algorithm for proposed attention-based word-level contextual feature extraction and cross-modality fusion

```
$Procedure\ FeatureExtraction()$                                        $Procedure\ to\ extract\ unimodal$
  $for\ i\ in\ 1\ to\ N\ do:$                                           $word\ level\ features$
    $for\ j\ in\ 1\ to\ M\ do:$                                         $where\ N\ is\ the\ number\ of\ utterances$
      $(t_{ij}, te_{ij}, td_{ij}) \longleftarrow TextFeatures(W_{ij})$     $and\ M\ is\ the\ number\ of\ words$
      $(a_{ij}, ae_{ij}, ad_{ij}) \longleftarrow AudioFeatures(W_{ij})$
      $(v_{ij}, ve_{ij}, vd_{ij}) \longleftarrow VideoFeatures(W_{ij})$

$Procedure\ BimodalFusion(p, q)$                                        $Procedure\ for\ Bimodal\ Fusion$
  $For\ i\ in\ 1\ to\ N\ do:$                                           $where\ p\ !=\ q \in \{t, a, v\}$
    $For\ j\ in\ 1\ to\ M\ do:$
      $P_{ij} = p_{ij}pd_{ij}$                                          $Word-level\ feature\ representation$
      $Q_{ij} = q_{ij}qd_{ij}$
      $PQ_{ij} = P_{ij}Q_{ij}$                                          $Concatenation\ of\ 2-2\ modalities$
      $pq\_s_{ij} = biLSTM(PQ_{ij})$                                    $Modality\ Specific\ Attention$
$Modality_{Score} = Importance(p, q)$                                   $Modality\ Score\ calculation$
$F_{pq} = Fusion\ (pq\_s,\ Modality_{Score})$                           $Cross-Modality\ Fusion$

$Procedure\ TrimodalFusion(t, a, v)$                                    $Procedure\ for\ Trimodal\ fusion$
  $For\ i\ in\ 1\ to\ N\ do:$
    $For\ j\ in\ 1\ to\ M\ do:$
      $T_{ij} = t_{ij}td_{ij}$                                          $Word-level\ featuree\ representation$
      $A_{ij} = a_{ij}ad_{ij}$
      $V_{ij} = v_{ij}vd_{ij}$
      $TAV_{ij} = T_{ij}A_{ij}V_{ij}$                                   $Concatenation\ of\ all\ modalities$
      $tav\_s_{ij} = biLSTM(TAV_{ij})$                                  $Modality\ Specific\ Attention$
$Modality_{Score} = Importance(t, a, v)$                                $Modality\ Score\ calculation$
$F_{tav} = Fusion\ (tav\_s,\ Modality_{Score})$                         $Cross-Modality\ Fusion$

$Procedure\ Classification\ (U)$                                        $Procedure\ for\ classification\ of$
  $for\ i\ in\ 1\ to\ N\ do:$                                           $utterance\ into\ discrete$
      $p\ (y|U) = softmax\ (\ w^{(s)}f_m + b^{(s)}\ )$                   $number\ of\ classes$
    $\hat{y} = \underset{y}{argmax}\ p\ (y|U)$
  $return\ (\hat{y})$
$FeatureExtraction()$                                                   $Unimodal\ Feature\ Extraction$
$F_{ta} \longleftarrow BimodalFusion(F_t, F_a)$                         $Bimodal\ Fusion$
$F_{tv} \longleftarrow BimodalFusion(F_t, F_v)$
$F_{av} \longleftarrow BimodalFusion(F_a, F_v)$
$F_{tav} \longleftarrow TrimodalFusion(F_{ta}, F_{tv}, F_{av})$         $Trimodal\ Fusion$
$C \longleftarrow Classification\ (U)$                                  $Classification$
```

The proposed attention-based word-level feature extraction and cross-modality fusion are shown in Figure 3 and the algorithm for the proposed methodology is summarised in Table 3.

4 Experimental results and discussion

The proposed word-level feature extraction and cross-modality fusion is implemented using Keras library with Tensorflow as backend. Firstly, the results of the proposed method are compared against the state-of-the-art methods (Poria et al., 2016; Zadeh et al., 2017). In Poria et al. (2016) textual features are extracted using convolutional neural networks, visual features using CLM-Z and acoustic features using OpenSmile open-source toolkit. Feature (early) level fusion technique was used to merge the affective information extracted from unimodal feature vectors. The merged trimodal feature vector is fed to multiple-kernel learning (MKL) based classifier. In Zadeh et al. (2017) pre-trained Glove embedding is used to extract textual features, acoustic features are extracted using COVAREP (Degottex et al., 2014) and facial expression analysis framework (FACET) (http://goo.gl/1rh1JN) is used to extract visual features. Novel tensor-based fusion technique was proposed and used for fusing information from multiple modalities. Extensive experimentation is conducted on two publically published datasets IEMOCAP (Section 3.1.1) and CMU-MOSI (Section 3.1.2).

Tables 4 and 5 shows the comparison of the performance of the proposed method with the baselines (Poria et al., 2016; Zadeh et al., 2017) in multimodal sentiment and emotion classification. The impact of recurrent units such as GRU, biGRU, biGRU with attention, LSTM, biLSTM, and biLSTM with attention mechanism on the performance of the proposed method is analysed. The results in Tables 4 and 5 indicate that the biLSTM with attention has performed better in combination modalities. Also, note that the proposed model outperforms baseline in all combination of modalities except audio and video. Tables 6 and 7 shows the confusion matrix for further analysis and Figures 4 and 5 shows the comparison of the performance of the proposed method with baselines for emotion and sentiment classification using IEMOCAP and CMU-MOSI dataset respectively.

Table 4 Comparison of performance of the proposed method with baselines for emotion classification using IEMOCAP dataset

			Proposed models with word-level features and cross-modality fusion					
Modality	*Poria et al. (2016)*	*Zadeh et al. (2017)*	*GRU*	*biGRU*	*biGRU with attention*	*LSTM*	*biLSTM*	*biLSTM with attention*
T + A	73.7%	71.1%	74.42%	75.24%	77.64%	76.32%	77.40%	79.55%
T + V	74.1%	73.7%	74.25%	75.16%	77.64%	76.24%	76.82%	79.63%
A + V	68.4%	67.4%	65.28%	65.60%	67.56%	66.56%	67.3%	68.56%
T + A + V	74.1%	73.6%	77.23%	77.48%	79.71%	78.72%	80.21%	81.62%

Legend: T: Text, A: Audio, V: Video, GRU: Gated recurrent unit, biGRU: Bidirectional GRU, LSTM: Long short term memory, biLSTM: Bidirectional LSTM.

Table 5 Comparison of performance of the proposed method with baselines for sentiment classification using CMU-MOSI dataset

			Proposed models with word-level features and cross-modality fusion					
Modality	*Poria et al. (2016)*	*Zadeh et al. (2017)*	*GRU*	*biGRU*	*biGRU with attention*	*LSTM*	*biLSTM*	*biLSTM with attention*
T + A	77.3%	77.0%	76.86%	77.39%	78.85%	78.05%	78.32%	79.25%
T + V	77.8%	77.1%	76.59%	77.79%	79.52%	78.22%	78.47%	79.38%
A + V	57.9%	56.5%	57.32%	57.64%	58.93%	57.64%	58.13%	59.83%
T + A + V	78.7%	77.2%	77.28%	78.18%	79.65%	79.12%	79.38%	80.45%

Figure 4 Comparison of experimental results on IEMOCAP dataset (see online version for colours)

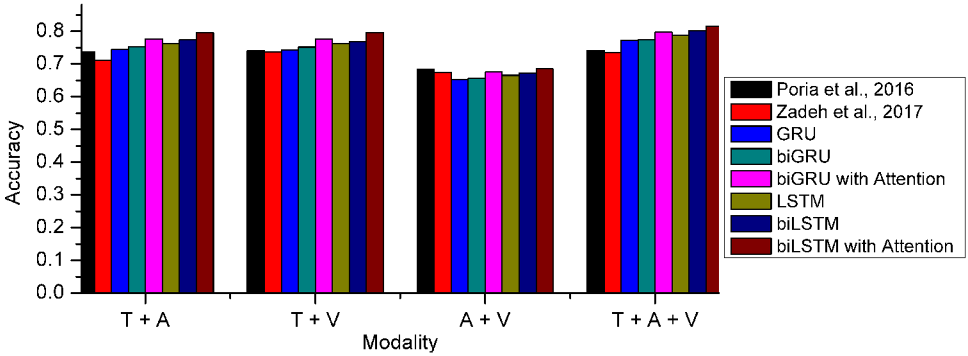


Figure 5 Comparison of experimental results on CMU-MOSI dataset (see online version for colours)

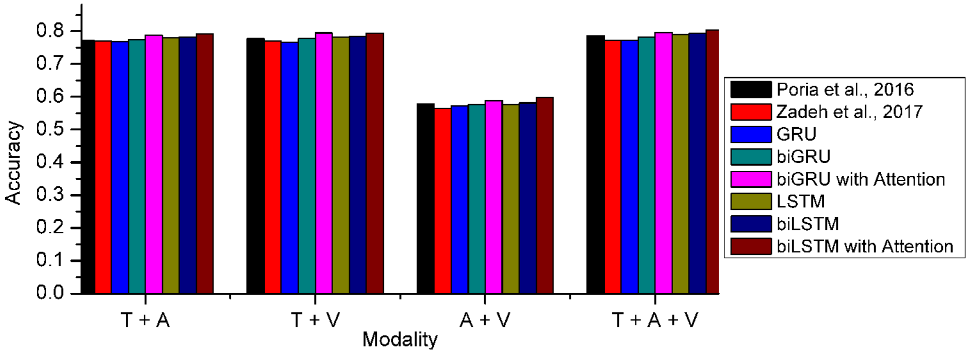


Table 6 Confusion matrix of the proposed method on IEMOCAP dataset (see online version for colours)

	Bimodal				*Bimodal*				*Bimodal*				*Trimodal*			
	Text – Audio				*Text – Video*				*Audio –Video*				*Audio – Video – Text*			
	H	*A*	*S*	*N*	*H*	*A*	*S*	*N*	*H*	*A*	*S*	*N*	*H*	*A*	*S*	*N*
H	89.15	5.08	5.77	0.00	79.68	3.46	16.86	0.00	79.45	4.85	11.78	1.62	86.84	1.62	11.55	0.00
A	2.94	73.11	19.33	4.62	2.94	68.49	24.79	3.78	1.68	64.71	22.27	11.34	1.26	77.31	19.33	2.10
S	13.16	1.84	73.68	11.32	13.95	3.16	75.26	7.63	13.68	2.11	59.21	25.00	11.32	1.84	80.26	6.58
N	0.00	3.82	19.11	77.07	0.00	1.27	13.38	85.35	0.00	1.27	10.19	88.54	0.00	4.46	18.47	77.07

Legend: H: Happy, A: Angry, S: Sadness and N: Neutral.

Table 7 Confusion matrix of the proposed method on CMU-MOSI dataset (see online version for colours)

	Bimodal		*Bimodal*		*Bimodal*		*Trimodal*	
	Text – Audio		*Text – Video*		*Audio – Video*		*Audio – Video – Text*	
	N	*P*	*N*	*P*	*N*	*P*	*N*	*P*
N	68.42	31.58	65.61	34.39	61.40	38.60	69.12	30.88
P	14.13	85.87	12.21	87.79	38.76	61.24	12.63	87.37

Legend: P: Positive and N: Negative.

5 Conclusion and future work

Multimodal fusion and the feature alignment among the modalities are the important challenges in multimodal sentiment classification and emotion recognition. To align the feature across the modalities, word-level features were extracted followed by bidirectional LSTM is used to extract the contextual information between the words of an utterance. The proposed method uses a modality-specific attention model to select the important word-level contextual features within an utterance of individual modality and modality attention is used to calculate modality score by understanding the importance of each modality before fusion. The proposed cross-modality fusion technique addresses the limitations of feature (early) level, model-based and decision (late) level fusion by fusing the important contextual features and modality score. The proposed model is tested on two standard publically available datasets IEMOCAP for emotion classification and CMU-MOSI for sentiment classification and the results show that the performance of the proposed method is better than the state of the art baselines in terms of classification accuracy. In the future, the work can be extended to select the class-specific feature and quality of unimodal features specifically audio modality to improve the performance in terms of classification accuracy.

References

Abanda, F.H. (2017) 'The application of controlled natural language on carbon market domain knowledge for enhanced retrieval of information', *International Journal of Intelligent Engineering Informatics*, Vol. 5, No. 4, pp.352–375.

Abdellaoui, M. and Douik, A. (2018) 'Template matching approach for automatic human body tracking in video', *International Journal of Intelligent Engineering Informatics*, Vol. 6, No. 5, pp.434–447.

Busso, C., Bulut, M., Lee, C-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S. (2008) 'IEMOCAP: interactive emotional dyadic motion capture database', *Language Resources and Evaluation*, Vol. 42, No. 4, pp.335–359.

Degottex, G., Kane, J., Drugman, T., Raitio, T. and Scherer, S. (2014) 'COVAREP-a collaborative voice analysis repository for speech technologies', *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp.960–964.

Du, P., Li, E., Xia, J., Samat, A. and Bai, X. (2018) 'Feature and model level fusion of pretrained CNN for remote sensing scene classification', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp.1–12.

Gohil, S., Vuik, S. and Darzi, A. (2018) 'Sentiment analysis of health care tweets: review of the methods used', *JMIR Public Health and Surveillance*, Vol. 4, No. 2, p.e43.

Gu, Y., Chen, S. and Marsic, I. (2018) 'Deep multimodal learning for emotion recognition in spoken language', *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp.5079–5083.

Huddar, M.G., Sannakki, S.S. and Rajpurohit, V.S. (2018) 'An ensemble approach to utterance level multimodal sentiment analysis', *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Belgaum, India, pp.145–150.

Huddar, M.G., Sannakki, S.S. and Rajpurohit, V.S. (2019a) 'A survey of computational approaches and challenges in multimodal sentiment analysis', *International Journal of Computer Sciences and Engineering*, Vol. 7, No. 1, pp.876–883.

Huddar, M.G., Sannakki, S.S. and Rajpurohit, V.S. (2019b) 'Multimodal emotion recognition using facial expressions, body gestures, speech, and text modalities', *International Journal of Engineering and Advanced Technology (IJEAT)*, Vol. 8, No. 5, pp.2453–2459.

Kirilenko, A.P., Stepchenkova, S.O., Kim, H. and Li, X. (2018) 'Automated sentiment analysis in tourism: comparison of approaches', *Journal of Travel Research*, Vol. 57, No. 8, pp.1012–1025.

Li, X., Xie, H., Chen, L., Wang, J. and Deng, X. (2014) 'News impact on stock price return via sentiment analysis', *Knowledge-Based Systems*, Vol. 69, pp.14–23.

Mahmoud, H., Abbas, E. and Fathy, I. (2018) 'Data mining and ontology-based techniques in healthcare management', *International Journal of Intelligent Engineering Informatics*, Vol. 6, No. 6, pp.509–526.

Mars, A. and Gouider, M.S. (2017) 'Big data analysis to features opinions extraction of customer', *Procedia Computer Science*, Vol. 112, pp.906–916.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) 'Distributed representations of words and phrases and their compositionality', *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol. 2, pp.3111–3119.

Mohammad, S.M., Kiritchenko, S. and Zhu, X. (2013) 'NRC-Canada: building the state-of-the-art in sentiment analysis of tweets', *Second Joint Conference on Lexical and Computational Semantics*, Canada, pp.321–327.

Nagamma, P., Pruthvi, H.R., Nisha, K.K. and Shwetha, N.H. (2015) 'An improved sentiment analysis of online movie reviews based on clustering for box-office prediction', *International Conference on Computing, Communication & Automation*, Noida, India, pp.933–937.

Pérez-Rosas, V., Mihalcea, R. and Morency, L.P. (2013) 'Utterance-level multimodal sentiment analysis', *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, August, Sofia, Bulgaria, pp.973–982.

Poria, S., Cambria, E. and Gelbukh, A. (2015) 'Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis', *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp.2539–2544.

Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017b) 'A review of affective computing: from unimodal analysis to multimodal fusion', *Information Fusion*, Vol. 37, pp.98–125.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A. and Morency, L.P. (2017a) 'Context-dependent sentiment analysis in user-generated videos', *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, pp.873–883.

Poria, S., Chaturvedi, I., Cambria, E. and Hussain, A. (2016) 'Convolutional MKL based multimodal emotion recognition and sentiment analysis', *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain, pp.439–448.

Ramteke, J., Shah, S., Godhia, D. and Shaikh, A. (2016) 'Election result prediction using Twitter sentiment analysis', *2016 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, pp.1–5.

Rosas, V.P., Mihalcea, R. and Morency, L-P. (2013) 'Multimodal sentiment analysis of Spanish online', *IEEE Intelligent Systems*, Vol. 28, No. 3, pp.38–45.

Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R. and Prasad, R. (2012) 'Ensemble of SVM trees for multimodal emotion recognition', *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Hollywood, CA, USA, pp.1–4.

Sakoe, H. and Chiba, S. (1978) 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 26, No. 1, pp.43–49.

Savran, A., Cao, H., Shah, M., Nenkova, A. and Verma, R. (2012) 'Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering', *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, Santa Monica, California, USA, pp.485–492.

Tao, F. and Liu, G.F. (2018) 'Advanced LSTM: a study about better time dependency modeling in emotion recognition', *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp.2906–2910.

Trabelsi, I., Bouhlel, M.S. and Dey, N. (2017) 'Discrete and continuous emotion recognition using sequence kernels', *International Journal of Intelligent Engineering Informatics*, Vol. 5, No. 3, pp.194–205.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K. and Morency, L-P. (2013) 'YouTube movie reviews: sentiment analysis in an audio-visual context', *IEEE Intelligent Systems*, Vol. 28, No. 3, pp.46–53.

Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L-P. (2017) 'Tensor fusion network for multimodal sentiment analysis', *Empirical Methods in Natural Language Processing, EMNLP*, Copenhagen, Denmark, 1103–1114.

Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E. and Morency, L-P. (2018) ‘Memory fusion network for multi-view sequential learning’, *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, Louisiana, USA, pp.5634–5641.

Zadeh, A., Zellers, R., Pincus, E. and Morency, L-P. (2016) ‘Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages’, *IEEE Intelligent Systems*, Vol. 31, No. 6, pp.82-88.

Website

http://goo.gl/1rh1JN