Information Fusion 76 (2021) 204–226

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey

Sarah A. Abdu [a,*], Ahmed H. Yousef [a,b], Ashraf Salem [a]

[a] Ain Shams University, Faculty of Engineering, Computers & Systems Department, Egypt
[b] Nile University, Center of Informatics Science (CIS), Egypt

ARTICLE INFO

Keywords:
Sentiment analysis
Sentiment classification
Multimodal sentiment analysis
Multimodal fusion
Audio, visual and text information fusion

ABSTRACT

Deep learning has emerged as a powerful machine learning technique to employ in multimodal sentiment analysis tasks. In the recent years, many deep learning models and various algorithms have been proposed in the field of multimodal sentiment analysis which urges the need to have survey papers that summarize the recent research trends and directions. This survey paper tackles a comprehensive overview of the latest updates in this field. We present a sophisticated categorization of thirty-five state-of-the-art models, which have recently been proposed in video sentiment analysis field, into eight categories based on the architecture used in each model. The effectiveness and efficiency of these models have been evaluated on the most two widely used datasets in the field, CMU-MOSI and CMU-MOSEI. After carrying out an intensive analysis of the results, we eventually conclude that the most powerful architecture in multimodal sentiment analysis task is the Multi-Modal Multi-Utterance based architecture, which exploits both the information from all modalities and the contextual information from the neighbouring utterances in a video in order to classify the target utterance. This architecture mainly consists of two modules whose order may vary from one model to another. The first module is the Context Extraction Module that is used to model the contextual relationship among the neighbouring utterances in the video and highlight which of the relevant contextual utterances are more important to predict the sentiment of the target one. In most recent models, this module is usually a bidirectional recurrent neural network based module. The second module is an Attention-Based Module that is responsible for fusing the three modalities (text, audio and video) and prioritizing only the important ones. Furthermore, this paper provides a brief summary of the most popular approaches that have been used to extract features from multimodal videos in addition to a comparative analysis between the most popular benchmark datasets in the field. We expect that these findings can help newcomers to have a panoramic view of the entire field and get quick experience from the provided helpful insights. This will guide them easily to the development of more effective models.

1. Introduction

Sentiments play a very important role in our daily lives. They help us to communicate, learn and make decisions, that's why over the past two decades, AI researchers have been trying to make machines capable of analyzing human sentiments. The early efforts in sentiment analysis have focused on textual sentiment analysis where only words are used to analyze the sentiment. However, textual sentiment analysis is insufficient to extract the sentiment expressed by humans; the meaning of words and sentences spoken by speakers often changes dynamically according to the non-verbal behaviours [1]. For example if someone said the word 'Amazing', it could express negative sentiment if it is accompanied by a sarcastic laugh or sarcastic voice.

Intensive research over many years have shown that multimodal systems are more efficient in recognizing the sentiment of a speaker than unimodal systems. The way humans naturally communicate and express their emotions and sentiments is usually multimodal: the textual, audio, and visual modalities are concurrently fused to extract information conveyed during communication in an effective way. A survey of multimodal sentiment analysis published in 2015 [2] reported that "multimodal systems were consistently (85% of systems) more accurate than their best unimodal counterparts, with an average improvement of 9.83%". Also, several surveys published later suggest that textual information is not sufficient for predicting human sentiments especially in

* Corresponding author
E-mail address: s.abdelaziz014@gmail.com (S.A. Abdu).

https://doi.org/10.1016/j.inffus.2021.06.003
Received 1 March 2021; Received in revised form 30 May 2021; Accepted 6 June 2021
Available online 12 June 2021

cases of sarcasm or ambiguity [3-6]. For example, it is impossible to recognize the sentiment of a sarcastic sentence “Great" as negative considering only the textual information. However, if the system can access the visual modality, it can easily detect the unpleasant gestures of the speaker and would classify it with the negative sentiment polarity. Similarly, acoustic features play important roles in the correctness of the system. In 2018 Poria et al. [7] also introduced an intuitive explanation of the improved performance in the multimodal scenario by visualizations of MOSI dataset [8] where both unimodal features and multimodal features are used to give information regarding the dataset distribution (see Fig. 1). For the textual modality only, comprehensive clustering can be seen with substantial overlap. However, this overlap is reduced in multimodal [7].

With the recent growth in social media platforms and the advances in technology, people started recording videos and uploading them on social media platforms like YouTube or Facebook to inform subscribers about their views. These videos may be product reviews, movie reviews, political debates, consulting, or their views about any random topic. A video provides a good source for extracting multi-modal information. In addition to the visual frames, it also provides information such as acoustic and textual representation of spoken language. This is what urged most AI researchers to direct their research towards multimodal sentiment analysis to leverage the varieties of (often distinct) information from multiple sources for building a more efficient system. There are many alternative ways to fuse information from different modalities, however selecting the best way is challenging [9].

Multimodal sentiment analysis focuses on modelling intramodal dynamics (View-specific dynamics) and intermodal dynamics (Cross-view dynamics) [10]. Intra-modality dynamics (View-specific dynamics) mean the interactions within a specific modality, independent of other modalities. For example the interactions between words in a given sentence. Intra-modality dynamics are particularly challenging for the language analysis since multimodal sentiment analysis is performed on spoken language. A spoken opinion such as “I think it was alright . . . Hmmm . . . let me think . . . yeah . . . no . . . ok yeah” almost never happens in written text. This volatile nature of spoken opinions, where proper language structure is often ignored, complicates sentiment analysis. On the other hand, inter-modality dynamics (Cross-view dynamics) refer to the interactions between the different modalities and are divided into synchronous and asynchronous categories. An example of synchronous cross-view dynamics is a smile and a positive word occurring simultaneously. And an example of asynchronous cross-view dynamics is the delayed occurrence of a laughter after the end of sentence. The main challenge in multimodal sentiment analysis is intra-modal representation and selecting the best approach to fuse features from different modalities [1, 11, 12].

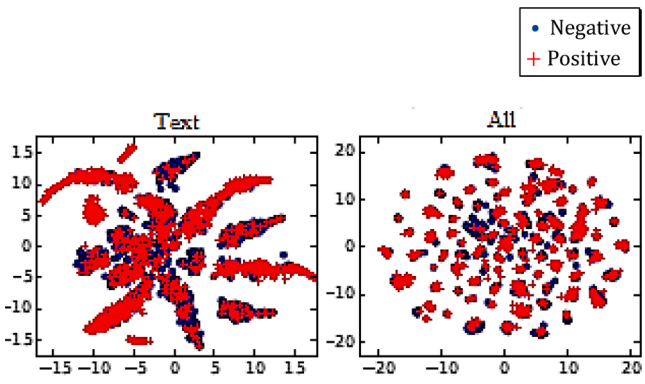

Fig. 1. T-SNE 2D visualization of MOSI dataset when text features and multimodal features are used [7].

1.1. The scope of this survey

As multimodal sentiment analysis research continues to gain popularity, the number of articles published every year in this field continues to increase which urges the need to have survey papers that summarize the recent research trends and directions in the field.

A long detailed survey was presented by *S. Poria et al.* [3] in which the authors presented an overview of the state of the art methodologies and trends in the field of multimodal sentiment analysis. However this survey was published in 2017 and many new articles and models have been introduced later, which urges the need for a new survey paper. For example, the most popular two datasets used in the field (CMU-MOSI and CMU-MOSEI) haven’t been summarized in [3]; this is mainly because the two datasets have been introduced after the paper was published. Furthermore, the most powerful models introduced in the field haven’t been referenced in this survey.

In 2021, *D. Gkoumas et al.* [13] replicated the implementation of eleven state of the art models in the field. They evaluated the performance of the referenced models for multimodal sentiment analysis tasks by using two benchmark datasets, CMU-MOSI [8] and CMU-MOSEI [14]. An experimental categorization of the models have also been provided with respect to the fusion approach used in each model, where the authors eventually conclude that the attention mechanism approaches are the most effective for the task. However, this paper has some shortcomings. First, *D. Gkoumas et al.* [13] completely ignored a certain category of models where the contextual information of neighboring utterances have been used to predict the sentiment of the target utterance, although these models have shown sophisticated performance in multi-modal sentiment analysis tasks. Second, the authors categorized only eleven models, where the categorization was based only on the fusion approach, while completely ignoring the architecture of each model. Third, they haven’t given any overview of the datasets used in the field except a brief summary of CMU-MOSI [8] and CMU-MOSEI [14].

The contribution of our survey is significant for many reasons. First, we noticed that each group of models have a common architecture, this is what urged us to categorize thirty-five models in the field into eight categories based on the architecture used in each model.

Second, we compare the effectiveness and efficiency of the thirty-five models on two widely used datasets for multimodal sentiment analysis (CMU-MOSI and CMU-MOSEI). After carrying out an intensive analysis of the results, we eventually come up to several conclusions. (1) We are able to conclude that the most powerful architecture in multimodal sentiment analysis task is the Multi-Modal Multi-Utterance based architecture, which exploits both the information from all modalities and the contextual information from the neighbouring utterances in a video in order to classify the target utterance. This architecture mainly consists of two modules whose order may vary from one model to another. The first module is the Context Extraction Module, which is used to model the contextual relationship among the neighbouring utterances in the video and highlight which of the relevant contextual utterances are more important to predict the sentiment of the target one. In most recent models, this module is usually a bidirectional recurrent neural network based module. The second module is the Attention-Based Module, which is responsible for fusing the three modalities (text, audio and video) and prioritizing only the important ones. (2) We were also able to conclude that using the scaled dot-product attention and the concept of multi-head attention are most effective for multimodal sentiment analysis task. (3) Moreover, the results obtained entailed that bimodal attention frameworks achieve better performance than self-attention frameworks. We believe that these findings could help the researchers to easily develop more effective models and choose the appropriate technique for a certain application.

Finally, in order to help new comers to have a panoramic view on the entire field, we provide brief details of the algorithms used in each model. In addition, we provide a comparative analysis between the most

popular benchmarks datasets in the field and a brief summary of the most popular approaches that have been used to extract features from multimodal videos.

In table 1, we present a brief comparison between the surveys presented by S. Poria et al. [3] and D. Gkoumas et al. [13], and our survey. The first column shows the reference to the survey while the second column shows the year when the survey was published. The third columns shows the number of models that have been categorized in each survey while the number of datasets summarized is given in the fourth column. The basis the authors used for categorizing the referenced models in their survey can be seen in the fifth column. The last column specifies whether the article reviewed the feature extraction methods by means of Yes/No answers (Y or N).

1.2. Multimodal sentiment analysis process on multimodal data

To classify the sentiment of any video, the visual, acoustic and textual features should be extracted first using the appropriate visual, acoustic and textual features extractors respectively. The extracted features of the three modalities are passed into a classification model to predict the correct sentiment. In the next sections, we will discuss different types of visual, acoustic and textual feature extractors and also tackle a comprehensive overview of the latest models used in the fields. The multimodal sentiment analysis process on multimodal data can be seen in Fig. 2.

This paper is organized as follows: Section 2 provides a summary of the most popular datasets in multimodal sentiment analysis. Section 3 tackles the most p feature extraction techniques and their related articles. Section 4 categorizes the state of the art models in multimodal sentiment analysis and the corresponding articles into eight architectures. Section 5 tackles the evaluation metrics used to evaluate the efficiency and effectiveness of the models. Section 6 presents the results and discussions, and finally the conclusion and future trend in research are tackled in Section 7.

2. Most popular datasets in multimodal sentiment analysis

The most popular datasets used in multimodal sentiment analysis are summarized in Table 2. The name of the dataset is shown in the first column while the year the dataset was published is presented in the second column. The number of videos in each dataset is shown on the third column while the number of utterances is shown in the fourth column. In some datasets the authors didn't mention anything about the number of utterances, that's why the forth column is sometimes left empty. The number of distinct speakers in the whole dataset can be seen in the fifth column. The sixth column shows the language in which the videos are recorded while the seventh column presents the source from which the videos were collected. The eighth column tells you from where you can download the dataset. The ninth column illustrates the sentiments that have been used for labeling the data. Finally the last column shows the topics of the videos included in the dataset, the videos could be product reviews, movie reviews, debates... etc.

2.1. YouTube Dataset

The YouTube Dataset was developed in 2011 by Morency et al. [15].

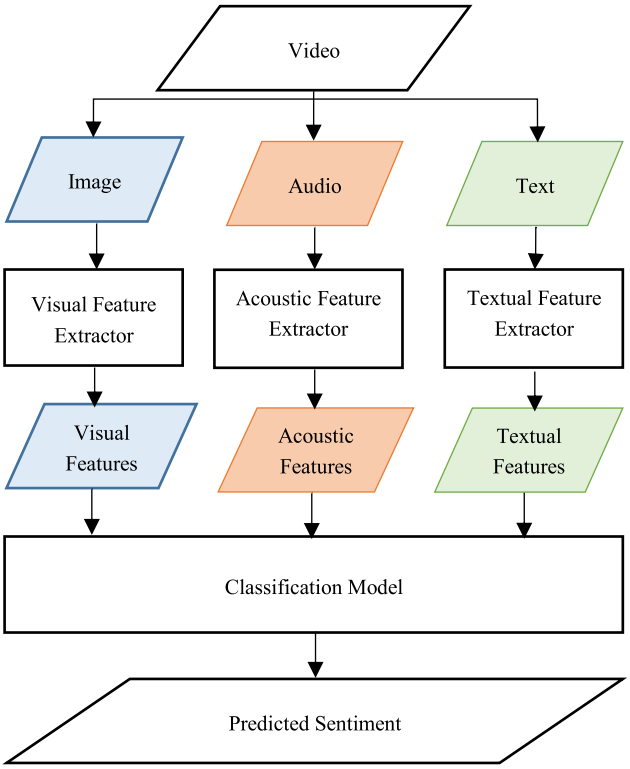

Fig. 2. Multimodal sentiment analysis process on a video.

The dataset was collected from YouTube website in such a way that it is not based on one particular topic where the videos were collected using the following keywords: opinion, review, product review, best perfume, toothpaste, war, job, business, cosmetics review, camera review, baby product review, I hate, I like... etc. Morency et al. [15] were so careful that the videos collected were diverse and having noise to encompass the different facets of sentiment analysis. The dataset consists of 47 generalized videos, each video contains 3-11 utterances. The ages of the speakers ranged from 14 to 60 years where 40 videos were expressed by female speakers, while the rest were expressed by male speakers. Although all speakers are from different cultures, they all expressed their opinions in English. Each video in the dataset was labelled with one of three sentiments: positive, negative or neutral giving a final set of 13 positively, 12 negatively and 22 neutrally labeled videos.

2.2. MOUD Dataset

The Multimodal Opinion Utterances Dataset (MOUD) was developed in 2013 by Perez-Rosas et al. [16] where 80 videos were collected from YouTube website using several keywords likely to lead to a product review or recommendation. The ages of the speakers ranged from 20 to 60 years where 15 videos were expressed by female speakers, while the rest were expressed by male speakers. All videos were recorded in Spanish. Eventually, a multimodal dataset of 498 utterances was created with an average duration of 5 seconds. Each utterance in the dataset was labelled with one of three sentiments: positive, negative or neutral

Table 1
Comparison between our survey and surveys introduced by *S. Poria et al.* [3] and *D. Gkoumas et al.* [13].

Paper	Year	# Models	# Datasets Summarized	Basis of Categorization	Feature Extraction methods overview
[3]	2017	>30	3	Fusion approach	Y
[13]	2020	11	2	Fusion approach	N
Ours	2021	35	7	Architecture of each model	Y

Table 2
Comparative analysis across most popular multimodal sentiment analysis datasets.

Dataset	Year	#V	#U	#S	Lang.	Source	Available at	Sentiments	Topics
YouTube	2011	47	280.	47	English	YouTube	By sending mail to stratou@ict.usc.edu	Positive, Negative, Neutral Sentiments	Product reviews
MOUD	2013	80	498	80	Spanish	YouTube	Publicly available http://web.eecs.umich.edu/mihalcea/downloads.html	Positive, Negative, Neutral	No particular topic
ICT-MMMO	2013	308	-	370	English	YouTube ExpoTV	By sending mail to stratou@ict.usc.edu	[-2,+2]	Movie Reviews
POM	2014	1000	-	352	English	ExpoTV	-	[1, 7]	Movie Reviews
CMU-MOSI	2016	93	2199	89	English	YouTube	Publicly available at http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/	[-3,+3]	No particular topic
CMU-MOSEI	2018	3228	22,777	1000	English	YouTube		[-3,+3]	Random topics, but the most frequent 3 topics are reviews (16.2%), debate (2.9%) and consulting (1.8%).
CH-SIMS	2020	60	2,281	474	Chinese	-	Publicly available https://github.com/thuiar/MMSA	[-2,+2]	Movies, TV series, and variety shows.

giving a final set of 182 positively, 231 negatively and 85 neutrally labeled videos.

2.3. ICT-MMMO Dataset

The Institute for Creative Technologies Multi-Modal Movie Opinion (ICT-MMMO) database was created in 2013 by *Wollmer et al.* [17]. The dataset consists of 370 online videos collected from YouTube and ExpoTV reviewing movies in English. Each video in the dataset was labelled with one of five sentiment labels: strongly positive, weakly positive, neutral, strongly negative and weakly negative.

2.4. PERSUASIVE OPINION MULTIMEDIA (POM) Dataset

S. Park et al. [18] collected 1000 movie reviews from ExpoTV where all reviewers expressed their opinions in English. Each movie review is a video of a speaker talking about a particular movie, as well as the speaker's direct rating of the movie on scale from 1 star (most negative review) to 5 stars (most positive review). Each video in the corpus has a frontal view of one person talking about a particular movie, and the average length of the videos is about 93 seconds. The dataset can be used for two purposes. First, it is used to study persuasiveness in the context of online social multimedia. Each video is annotated from 1 (very unpersuasive) to 7 (very persuasive). Second, it is used recognize the speaker traits. Each movie review video was annotated with one of the following speaker traits: confidence, entertaining, trusting, passion, relaxed, persuasive, dominance, nervous, credibility, entertaining, reserved, trusting, relaxed, nervous, humorous and persuasive. 903 videos were split into 600 for training, 100 for validation and 203 for testing.

2.5. CMU-MOSI Dataset

The CMU-MOSI Dataset was developed in 2016 by Amir Zadeh et al [8]. The dataset consists of 93 videos collected from YouTube video-blogs, or vlogs. The vblogs are YouTube videos where many users can express their opinions about many different subjects; this type of videos usually contain only one speaker, looking primarily at the camera.

The ages of the speakers ranged from 20 to 30 years where 41 videos were expressed by female speakers, while the rest were expressed by male speakers. Although all speakers are from different cultures, they all expressed their opinions in English. One big advantage of these videos is that they address diversity and contain noise; all videos were recorded in different setups, some users have high-tech microphones and cameras, while others use less professional recording devices. Also users are in different distances from the camera, and background and lighting conditions differed from one video to another. The videos were kept in their original resolution without any enhancement to the quality.

Each utterance in the dataset was labelled with one of the following sentiments: strongly positive (labelled as +3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), strongly negative (-3).

2.6. CMU-MOSEI Dataset

The CMU-MOSEI dataset was developed in 2018 by Zadeh et al [14]. CMU-MOSEI is a larger scale dataset that consists of 3228 videos divided into 22,777 utterances from more than 1000 online YouTube speakers (57% male to 43% female). The videos talk about 250 distinct topics but the most frequent 3 topics are reviews (16.2%), debate (2.9%) and consulting (1.8%). Each video contains only one speaker looking primarily at the camera. As CMU-MOSI, CMU-MOSEI also address diversity and contain noise. Each utterance in the dataset was labelled with one of eight sentiments: strongly positive (labeled as +3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), strongly negative (-3).

It is worth mentioning that recently most researches use CMU-MOSI and CMU-MOSEI datasets to evaluate the performance of their models in multimodal sentiment analysis.

2.7. CH-SIMS Dataset

The CH-SIMS Dataset was developed in 2020 by Yu et al. [19]. The dataset consists of 60 video divided into 2281 utterances collected from movies, TV series, and variety shows. Average length of the utterance is 3.67 seconds and in each video clip, no other faces appear except the face of the speaker. For each utterance, the annotators give one multimodal annotation and three unimodal annotations for each video clip. This can help researchers to use SIMS to do both unimodal and multimodal sentiment analysis tasks. Furthermore, researchers can develop new methods for multimodal sentiment analysis with these additional annotations. The annotations in this dataset can be one of the following: positive, weakly positive, neutral, weakly negative, negative.

3. Feature Extraction

3.1. Visual Feature Extraction

Facial expressions have always been the primary keys for analyzing the emotions and the sentiments of a speaker. Many measurement systems for facial expressions have been developed. The most famous one is the Facial Action Coding System (FACS) which has been developed by Ekman and Friesen and published in 1978. FACS is an anatomically based system for describing all visually discriminative facial movement. FACS is based on the reconstruction of facial expressions in terms of Action Units (AUs). The facial muscles of all humans are almost identical and AUs are based on movements of these muscles. FACS distinguishes facial actions only but doesn't identify the emotions. FACS codes are used to infer emotions using various resources available. Some resources

used a combinations of AUs to infer emotions such as FACS Investigators' Guide [20], the FACS interpretive database, and a large body of empirical research [21].

Ekman's work encouraged many researchers to exploit image and video processing methods in order to analyze facial expressions. Yacoob et al. [22] and Black et al. [23] used high gradient points on the face, and head and facial movements to recognize facial expressions. In 1999, Zhang et al. [24] used geometrical features with a multi-scale, multi--orientation Gabor Wavelet-based representation to identify expressions. In 2000, Haro et al. [25] used Kalman Filter and principal component analysis (PCA) to enhance the features. A stochastic gradient descent based technique [26] and active appearance model (AAM) [27] were used to recover the face shape and texture parameters, for facial features. Donato et al. [28] also provided a comparison of several techniques, such as optical flow, PCA, independent component analysis (ICA), local feature analysis and Gabor wavelet, for recognition of action units, and observed that, Gabor wavelet representation and ICA performed better on most datasets. In 2001, Tian et al. [29] claimed that every part of the face is an important feature, thus introduced a multi-state face component model to make use of both permanent and transient features. Permanent features are those which remain the same through ages, for example opening and closing of lips and eyes, pupil location, eyebrows and cheek areas. Transient features are observed only at the time of facial expressions, such as contraction of the corrugator muscle that produces vertical furrows between the eyebrows. Texture features of the face have also been used for facial expression analysis in a number of feature extraction methods, including: image intensity [30], image difference [31], edge detection [29], and Gabor wavelets [32]. In 2002, Ekman et al. [20] introduced an updated version of FACS where the description of each AU, and AU combinations were refined. Moreover, details on head movements and eye positions were added.

Many facial expression recognition techniques, face tracking methods and feature extraction methods have been introduced. The most popular ones are Active Appearance Models (AAM) [33], Optical flow models [22], Active Shape Models (ASM) [34], 3D Morphable Models (3DMM) [35], Muscle-based models [36], 3D wireframe models [37], Elastic net model [38], Geometry-based shape models [39], 3D Constrained Local Model (CLM-Z) [66], Adaptive View-based Appearance Model (GAVAM) [40].

All the pre-mentioned methods do not work well for videos because they do not model temporal information. An important facet in video-based methods is maintaining accurate tracking throughout the video sequence. A wide range of deformable models, such as muscle-based models [36], 3D wire frame models [37], elastic net models [38] and geometry-based shape models [39, 41], have been used to track facial features in videos. Following this, many automatic image-based and video-based methods for detection of facial features and facial expressions were proposed [42, 43].

Research in psychology proved that the body gestures can provide a great significance to the emotion and sentiment of the speaker. A detailed study was carried out to prove that body gestures are highly related to emotions of the speaker and that different emotions can result in various combinations of body gesture dimensions and qualities of the speaker [44]. This what urged some researchers to focus on extracting features from the body gestures for sentiment analysis and emotion recognition [45-48].

The previous sections described how to extract handcrafted features from a visual modality and how to create mathematical models for facial expression analysis. With the advent of deep learning, deep learning models can learn the best features automatically without prior intervention. The deep learning framework enables robust and accurate feature learning in both supervised and unsupervised settings, which in turn produces best performance on a range of applications, including digit recognition, image classification, feature learning, visual recognition, musical signal processing and NLP [3] . Motivated by the recent success of deep learning in feature extraction, sentiment analysis tasks started to adopt deep learning algorithms to extract visual features, especially the convolutional neural network (CNN). In [49], a novel visual sentiment prediction framework was designed to understand images using CNN. The framework is based on transfer learning from a CNN pre-trained on large scale data for object recognition, which in turn is used for sentiment prediction. The main advantage of the proposed framework is that there is no requirement of domain knowledge for visual sentiment prediction. In 2014 You et al. [50] employed 2D-CNN in visual sentiment analysis, coupled with a progressive strategy to fine tune deep learning networks to filter out noisy training data. They also used domain transfer learning to improve the performance. In 2015, Tran et al. [51] proposed a deep 3D convolutional network (3D-CNN) for spatiotemporal feature extraction. This network consists of 8 convolution layers, 5 pooling layers, 2 fully connected layers, and a softmax output layer. The network has proved to be very robust for extracting spatiotemporal features. In 2016 Poria et al. [52] proposed a convolutional recurrent neural network to extract visual features from multimodal sentiment analysis and emotion recognition datasets where a CNN and RNN have been stacked and trained together.

Recent trend in visual features extraction

For the past few years, many sentiment analysis models have used either 3D-CNN to extract the visual features from videos in the past few years, or used publicly available libraries to extract visual features like FACET, OKAO, and CERT libraries. FACET library is used to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features. OKAO Vision is a commercial software that detects the face at each frame, then it extracts the facial features and extrapolates some basic facial expressions as well as eye gaze direction. The main facial expression being recognized is smile. This is a well-established technology that can be found in many digital cameras. The Computer Expression Recognition Toolbox (CERT) [53] automatically extract the smile and head pose estimates, Facial AUs. These features describe the presence of two or more AUs that define one of eight emotions (anger, contempt, disgust, fear, joy, sad, surprise, and neutral). For example, the unit A12 describes the pulling of lip corners movement, which usually suggests a smile but when associated with a check raiser movement (unit A6), defines happiness emotion.

3.2. Acoustic Feature Extraction

Early research on audio features extraction focused on the acoustic properties of spoken language. Some psychological studies related to emotion showed that vocal parameters, especially pitch, intensity, speaking rate and voice quality have a great role in sentiment analysis and emotion recognition [54]. Some other studies showed that acoustic parameters change through both oral variations and personality traits. Many researches have been conducted to find the type of features that can be used for better analysis [55, 56], where researchers have found that pitch and energy related features play a key role in emotion recognition and sentiment analysis. Some other researchers used other features for feature extraction like pitch, pause duration, spectral centroid, spectral flux, beat histogram, beat sum, strongest beat, formants frequencies, mel-frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), log frequency power coefficients (LFPC), pause duration, teager energy operated based features. Please refer to [3] for brief description of these features.

It is also worth mentioning that some articles have classified the affective reactions to sound into discrete feeling states and states based on dimensions [57, 58]. Discrete feeling states are defined as spontaneous, uncontrollable emotions. The states based on dimension are hedonic valence (pleasantness), arousal (activation, intensity) and dominance.

Audio affect classification has also been classified into local features

and global features. The common approach to analyze audio modality is to segment each utterance into either overlapped or non-overlapped segments and examine them. Within a segment the signal is considered to be stationary. The features extracted from these segments are called local features. In speech production, there are several utterances and, for each utterance, the audio signal can be divided into several segments. Global features are calculated by measuring several statistics such as the average, mean, deviation of the local features. Global features are the most commonly used features in the literature. They are fast to compute and, as they are fewer in number compared to local features, so the overall speed of computation is enhanced [59]. However, there are some drawbacks of calculating global features, as some of them are only useful to detect the effect of high arousal, e.g., anger and disgust. For lower arousals, global features are not that effective, e.g., global features are less prominent to distinguish between anger and joy. Global features also lack temporal information and dependence between two segments in an utterance.

The benchmark results proposed by Navas et al. [60] proved that the speaker-dependent approaches often gives much better results than the speaker-independent approaches. However, the speaker-dependent approach is not feasible in many practical applications that deal with a very large number of users.

As for computer vision, deep learning is increasingly gaining attention in audio classification research. A possible research question is whether deep neural networks can be replicated for automatic feature extraction from audible data. The answer to this question was given in 2015 by a group of researchers [61], where a CNN was used to extract features from audio, which were then used in a classifier for the final emotion classification task. Deep neural networks based on Generalized Discriminant Analysis (GerDA) are also a very popular approach in the literature to extract features automatically from raw audio data. However, most deep learning approaches in audio emotion classification literature rely on handcrafted features [62].

Recent trend in acoustic features extraction

Recently most multimodal sentiment analysis models use OpenSMILE [63], COVAREP [64], Open EAR [65] to extract acoustic features. They are freely available popular audio feature extraction toolkits which are able to extract all the key features as elaborated above. OpenSMILE feature extraction toolkit unites feature extraction algorithms from the speech processing and the Music Information Retrieval communities. It used to extract CHROMA and CENS features, loudness, MFCC, perceptual linear predictive cepstral coefficients, linear predictive coefficients, line spectral frequencies, fundamental frequency, and formant frequencies [63]. COVAREP is used to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch, and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients. All extracted features are related to emotions and tone of speech. The audio features are automatically extracted from the audio track of each video clip [64]. OpenEAR is an open source software that can be used to automatically compute the pitch and voice intensity. Speaker normalization is performed using z-standardization [65].

3.3. Textual Feature Extraction

Traditionally, the bag-of-words (BoW) model had been used to extract features for sentences and documents in NLP and text mining. It is called a "*bag*" of words, because any information about the order or structure of words in the document is ignored. The model is only concerned with whether known words occur in the document, not where in the document.

Based on BoW, a document is transformed to a numeric feature vector with a fixed length, where each element in the vector is scored. This score can be: a binary scoring of the presence or absence of words, word frequency, or TF–IDF score. Despite its popularity, BoW has some shortcomings. First, the dimension of this vector is equal to the size of the vocabulary, so as the vocabulary size increases, the vector representation of documents increases also. You can imagine that for a very large corpus, such as thousands of books, that the length of the vector might be thousands or millions of positions. Second, BoW can barely encode the semantics of words since the word order is ignored, which means that two documents can have exactly the same representation as long as they share the same words. Third, each document may contain very few number of the known words in the vocabulary which results in a vector with lots of zero scores, called a sparse vector or sparse representation. Sparse vectors require more memory and computational resources when modeling and the vast number of positions or dimensions can make the modeling process very challenging for traditional algorithms.

Later, a more sophisticated model was introduced to create a vocabulary of grouped words called Bag-of-n-grams, an extension for BoW. This changes the scope of the vocabulary and allows the bag-of-words to capture a little bit more meaning from the document. In this approach, each word or token is called a "gram". Creating a vocabulary of two-word pairs is, in turn, called a bigram model. Again, only the bigrams that appear in the corpus are modeled, not all possible bigrams. In this mode, the scores can highlight words that are distinct (contain useful information) in a given document. Thus, this model can consider the word order in a short context (n-gram), however it still suffers from data sparsity and high dimensionality [66].

To overcome the shortcomings of BoW and n-grams, word embedding techniques were proposed. A word embedding is a technique for feature extraction that uses neural networks to learn a representation for text such that words that have the same meaning have a similar representation. Word embedding transforms words in a vocabulary to vectors of continuous real numbers. The technique normally involves embedding high-dimensional sparse vector (e.g., one-hot vector) to a lower-dimensional dense vector which can encode some semantic and syntactic properties of words. Each dimension of the embedding vector represents a latent feature of a word. Word Embeddings solve the shortcomings of one hot vector and achieve dimensionality reduction. That's why it may be considered one of the key breakthroughs of deep learning on challenging natural language processing problems [66].

Recent trend in textual features extraction

More recently, the new trend in multimodal sentiment analysis focuses on using word embeddings pre-trained on a large corpus such as the Glove [67] or word2vec [68]. And all the models referenced in this paper use word embeddings to learn a representation for written text.

In Table 3 and Table 4, we present the methods used by each of the models referenced in this paper to extract visual, acoustic and textual features.

4. State of the art models in multimodal sentiment analysis

In this section, we classify thirty five recent models in multimodal sentiment analysis into eight categories, based on the architecture of each model as shown in Fig. 3.

4.1. Early Fusion based models (Feature Level Fusion)

In this category, all modalities are concatenated into a single view. Then this concatenated view is used as input to a prediction model as shown in Fig. 4a. The prediction models could be as simple as Hidden Markov Models (HMMs) [69], Support Vector Machines (SVMs) [70] or Hidden Conditional Random Fields (HCRFs) [71]. Later, after the advances of deep learning, Recurrent Neural Networks, especially Long-short Term Memory [72] have been used for sequence modeling. Although this simple concatenation succeeded somehow in modeling

Table 3
Effectiveness results for some multimodal sentiment analysis models on MOSI dataset and the feature extraction methods used in each model.

		Efficiency					Feature Extraction		
		Binary Classification		Multiclass	Regression		Textual	Visual	Acoustic
Ref	Model	Acc_2	F1 Score	Acc7	MAE	Corr			
[74]	THMM	50.7%	45.4%	17.8%	-	-	Glove	FACET	COVAREP
[74]	SVM	71.6%	72.3%	26.5%	1.1	0.559	Glove	FACET	COVAREP
[13]	EF-LSTM	75.8%	75.6%	32.7-%	1.000	0.630	Glove	FACET	COVAREP
[13]	LF-LSTM	76.2%	76.2%	32.7%	0.987	0.624	Glove	FACET	COVAREP
[74]	DF	72.3%	72.1%	26.8%	1.143	0.518	Glove	FACET	COVAREP
[74]	Majority	50.2 %	50.1%	17.5%	1.864	0.057	Glove	FACET	COVAREP
[74]	MARN	77.1%	77%	34.7%	0.968	0.625	Glove	FACET	COVAREP
[73]	MFN	77.4%	77.3%	34.1%	0.965	0.632	Glove	FACET	COVAREP
[80]	RMFN	78.4%	78%	38.3%	0.922	0.681	Glove	FACET	COVAREP
[89]	Multilogue-Net	81.19%	80.10%	-	-	-	CNN	3D-CNN	openSMILE
[74]	TFN	74.6%	74.5%	28.7%	1.040	0.587	Glove	FACET	COVAREP
[81]	LMF	76.4%	75.7%	32.8%	0.912	0.668	Glove	FACET	COVAREP
[84]	MRRF	77.46%	76.73%	33.02%	0.912	0.772	Glove	FACET	COVAREP
[83]	HFFN	80.19%	80.34%	-	-	-	Glove	FACET	COVAREP
[85]	MMUUSA	79.52	-	-	-	-	word2vec	3D-CNN	openSMILE
[13]	MMUUBA	78.2%	78.1%	33.8%	0.947	0.675	Glove	FACET	COVAREP
[86]	BIMHA	76.68%	76.6%	36.15%	0.9694	0.644	BERT [105]	LibROSA [106]	LibROSA
[87]	SWAFN	80.2%	80.1%	40.1%	0.88	0.697	GloVe	FACET	COVAREP
[13]	RAVEN	78.6%	78.6%	34.6%	0.948	0.674	Glove	FACET	COVAREP
[90]	Bc-LSTM	80.3%	-	-	-	-	word2vec	3D-CNN	openSMILE
[85]	MMMU-BA	82.31%	-	-	-	-	word2vec	3D-CNN	openSMILE
[94]	MMMUBA 2	80.33%	-	-	-	-	word2vec	3D-CNN	openSMILE
[1]	MHSAN	78.7%	-	-	-	-	word2vec	3D-CNN	openSMILE
[98]	MHAM	82.71%	-	-	-	-	-	-	-
[95]	Bi-LSTM	81.3%	-	-	-	-	word2vec	3D-CNN	openSMILE
[96]	MARNN	84.31%	-	-	-	-	word2vec	3D-CNN	openSMILE
[13] [99]	MCTN	79.3%	79.1%	32.3%	0.909	0.676	-	-	-
[100]	MulT	83.0%	82.8%	40.0%	0.871	0.698	Glove	FACET	COVAREP
[103]	QMF	80.69%	79.77%	47.88%	0.6399	0.6575	-	-	-

multi-view problems, it causes over-fitting in case of a small size training dataset and is not intuitively meaningful because modeling view-specific dynamics is ignored, thus losing the context and temporal dependencies within each modality [73]. Some of the models that used this architecture are reviewed in this section.

4.1.1. THMM (Tri-modal Hidden Markov Model)

Morency at al. [15] were the first to address the problem of tri-modal sentiment analysis. They used an HMM for classification after concatenation.

4.1.2. SVM (Support Vector Machine)

Perez-Rosas, Mihalcea, and Morency [16] combined the multimodal streams into a single feature vector, thus resulting in one vector for each utterance in the dataset, which is used by a SVM for binary classification to make a decision about the sentiment of the utterance. On the other hand *S. Park et al.* [18] used the Support Vector Machines (SVMs) for classification and Support Vector Regression (SVRs) for regression experiments with the radial basis function kernel as the prediction models.

4.1.3. EF-LSTM (Early Fusion LSTM)

Zadeh et al. [74] and *D. Gkoumas et al.* [13] concatenate the different

Table 4
Effectiveness results for some multimodal sentiment analysis models on MOSEI dataset and the feature extraction methods used in each model.

		Efficiency					Feature Extraction		
		Binary Classification		Multiclass	Regression		Textual	Visual	Acoustic
Ref	Model	Acc_2	F1 Score	Acc7	MAE	Corr			
[13]	EF-LSTM	78.2%	77.1%	45.7%	0.687	0.573	Glove	FACET	COVAREP
[13]	LF-LSTM	79.2%	78.5%	47.1%	0.655	0.614	Glove	FACET	COVAREP
[74]	DF	72.3%	72.1%	26.8%	1.143	0.518	Glove	FACET	COVAREP
[74]	Majority	50.2 %	50.1%	17.5%	1.864	0.057	Glove	FACET	COVAREP
[13]	MFN	79.9%	79.1%	47.4%	0.646	0.626	Glove	FACET	COVAREP
[14]	Graph MFN	76.9%	77.0%	45.0%	0.71	0.54	-	-	-
[13]	MARN	79.3%	77.8%	47.7%	0.646	0.629	Glove	FACET	COVAREP
[89]	Multilogue-Net	82.10%	80.01%	-	0.59	0.5	Glove	FACET	OpenSMILE
[13]	TFN	75.6%	75.5%	34.9%	1.009	0.605	Glove	FACET	COVAREP
[13]	LMF	78.2%	77.6%	47.6%	0.660	0.623	Glove	FACET	COVAREP
[85]	MMUUSA	79.76%	-	-	-	-	Glove	FACET	COVAREP
[13]	MMUUBA	80.7%	80.2%	48.4%	0.627	0.672	Glove	FACET	COVAREP
[13]	RAVEN	80.2%	79.8%	47.8%	0.636	0.654	Glove	FACET	COVAREP
[85]	MMMU-BA	79.8%	-	-	-	-	Glove	FACET	COVAREP
[100]	MulT	82.5%	82.3%	51.8%	0.580	0.703	Glove	FACET	COVAREP
[103]	QMF	79.74%	79.62%	33.53%	0.9146	0.6959	-	-	-

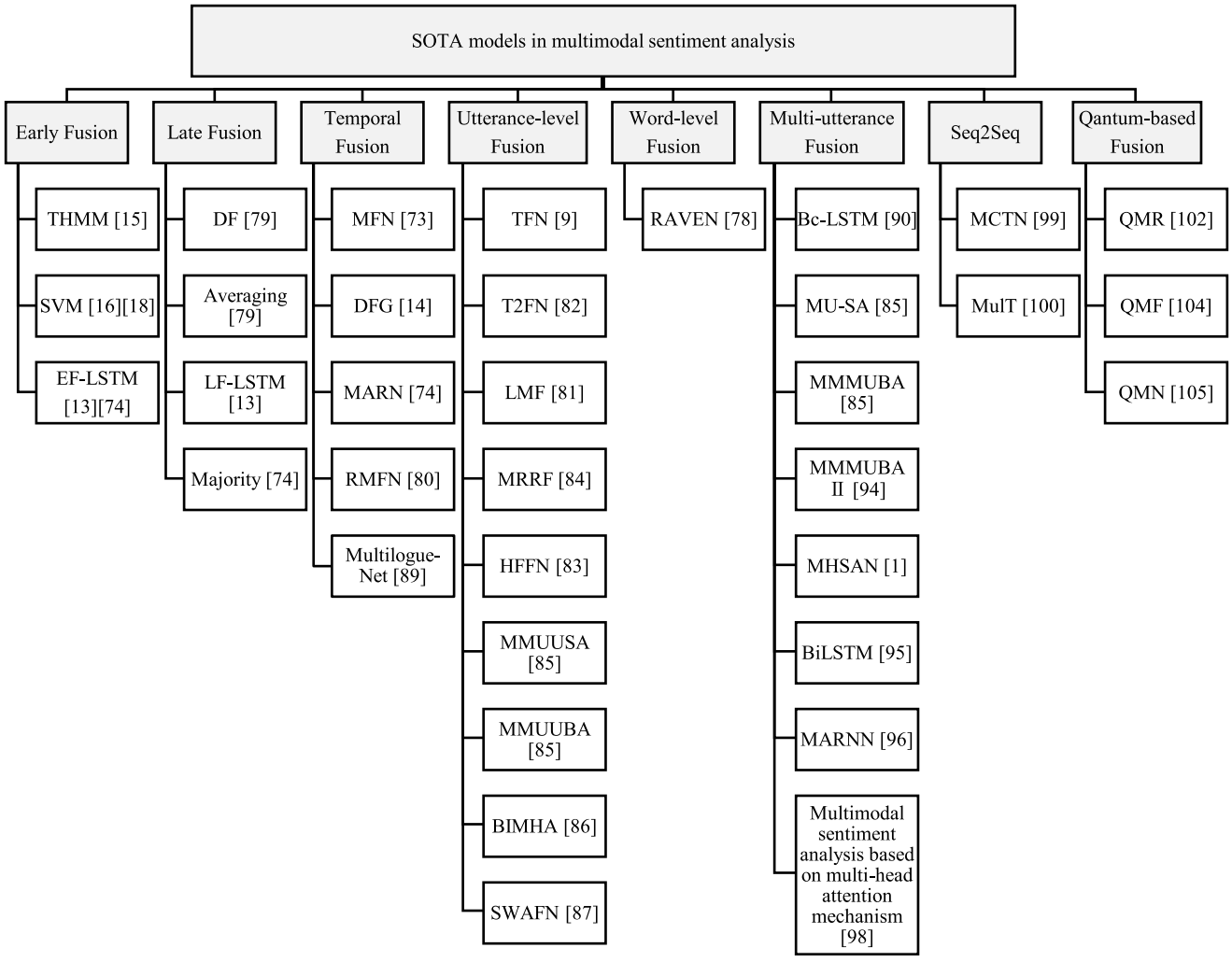

Fig. 3. SOTA models in multimodal sentiment analysis.

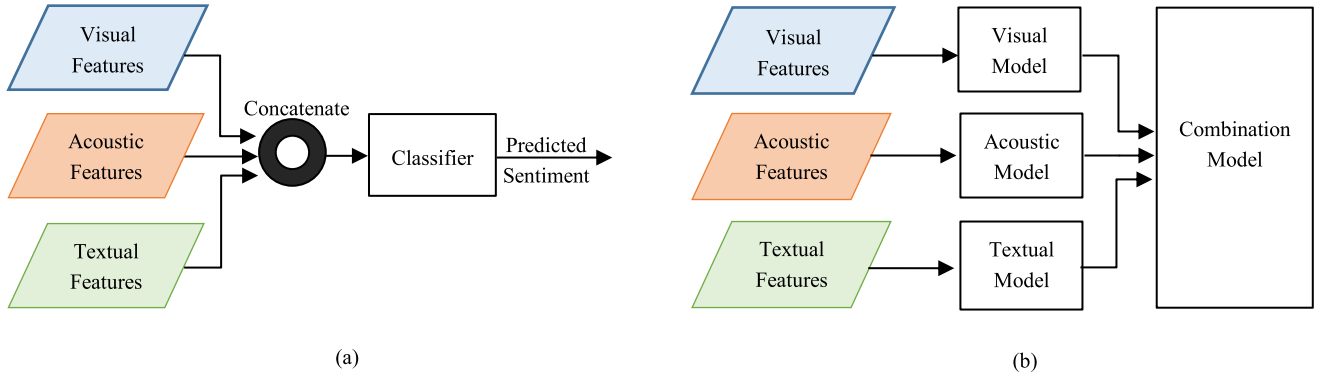

Fig. 4. Traditional fusion techniques (a) Early fusion architecture, (b) Late fusion architecture.

modalities at each timestep in a single feature vector and use it as input to an LSTM [75]. *Zadeh et al.* [74] used bidirectional stacked LSTM while *D. Gkoumas et al.* [13] passed the last hidden state of the LSTM to two fully connected layers to produce the output sentiment. Both reported a very slight difference in performance metrics.

4.2. Late Fusion based models

In late fusion, different models are built for each modality and then their decisions are combined by averaging, weighted sum [76], majority voting [77], or deep neural networks as shown in Fig. 4b. The late fusion method integrates different modalities at the prediction level.

The advantage of these models is that they are very modular, and

that one can build a multimodal model from individual pre-trained unimodal models with just fine-tuning on the output layer. These methods are generally strong in modeling view-specific dynamics and can also outperform unimodal models but they have problems in modelling cross-view dynamics since are normally more complex than a decision vote [78]. Some of the models that used this architecture are reviewed in this section.

4.2.1. Decision Voting (Deep Fusion DF)

Nojavanasghari et al. [79] train one deep neural model for each modality and performs decision voting on the output of each modality network.

4.2.2. Averaging

Nojavanasghari et al. [79] train one deep neural model for each modality. The output scores from all deep models are averaged.

4.2.3. LF-LSTM

D. Gkoumas et al. [13] builds separate LSTMs for textual, visual, and acoustic modalities, and then concatenates the last hidden state of the three LSTMs. The concatenated hidden states are passed into two fully connected layers to produce the output sentiment.

4.2.4. Majority

*Za*deh et al. [74] performs majority voting for classification tasks, and predicts the expected label for regression tasks.

4.3. Temporal based Fusion

In this architecture, the model accounts for view-specific and cross-view interactions and continuously models them over time with a special attention mechanism. In most models the architecture consists of two main components as shown in Fig. 5.

■System of LSTMs/LSTHMs

Each modality is assigned one LSTM or LSTHM to model its view-specific dynamic. At each time step time t, the modality specific features are fed into the corresponding LSTM/LSTHM.

- Attention Block

This layer is an explicitly designed attention mechanism responsible for attending to the most important components of the output of the systems of LSTMs/LSTHMs to model the cross-view dynamics.

All the models in this section differ only in the way they apply attention on the output of the LSTMs/LSTHMs system except Multilogue-Net whose architecture is somehow different. We will use the symbols x_a^t, x_v^t, x_ℓ^t to denote acoustic, visual and textual features respectively at time step (t).

4.3.1. Memory Fusion network (MFN)

At each timestamp (t) of MFN recursion, the memory of the three LSTMs are concatenated together with the memory of the three LSTMs at the previous timestamp $(t-1)$ and passed to the attention block. The attention block in this component is a simple neural network with softmax at the output layer to calculate the attention weights, and is called Dynamic Memory Attention Block (DMAN). The output of this module is the attended memories of the LSTMs, which is passed to a Multi-view Gated Memory [73]. The Multi-view Gated Memory is a unifying memory which stores the cross-view interactions over time. It has two gates, controlled by two neural networks, called retain and update respectively. At each timestep, the retain gate determine how much of the current state of the Multi-view Gated Memory to remember while the update gate determines how much of the Multi-view Gated Memory to update respectively as shown in Fig. 6.

Finally, the final state of the Multi-view Gated Memory and the output of the three LSTMs at the last timestamp of the input sequence are concatenated together to construct the multimodal sequence representation that can be used to produce the output sentiment using two fully connected layers [73].

The power of MFN is that DMAN can model asynchronous cross-view interactions because it attends to the memories in the System of LSTMs which can carry information about the observed inputs across different timestamps [73].

4.3.2. Graph MFN (DFG)

Zadeh et al. [14] replicate the architecture of MFN except that they replace DMAN in MFN [73] with a new neural-based component called the Dynamic Fusion Graph (DFG. Please refer to the original paper [14] for detailed explanation of DFG.

4.3.3. Multi-Attention Recurrent Network (MARN)

The architecture of MARN is very similar to MFN; the main difference is that Zadeh et al. [74] modified the LSTM to create a hybrid LSTM (LSTHM) where they reformulate the memory component of each LSTM to carry hybrid information; the view specific dynamics of its modality and the cross-view dynamics code related to that modality. The asynchronous cross-view dynamics are captured at each time-step using neural-based attention block called multi-attention block (MAB) as shown in Fig. 7. The authors claimed that there may exist more than one cross-view dynamics that occur simultaneously across the three modalities, thus MAB consists of (K) neural networks, each with softmax at the output layer, responsible for modelling (K) cross view dynamics. The

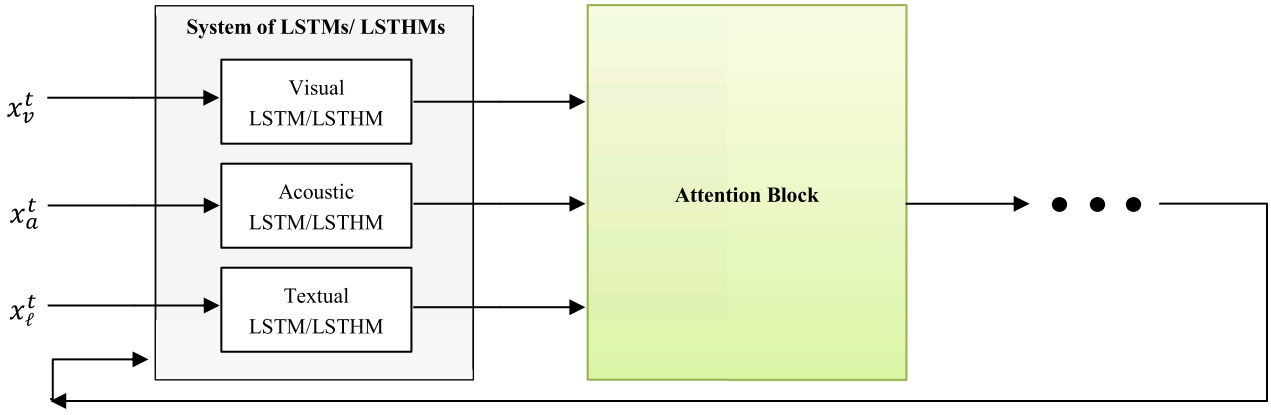

Fig. 5. Architecture of temporal fusion based models.

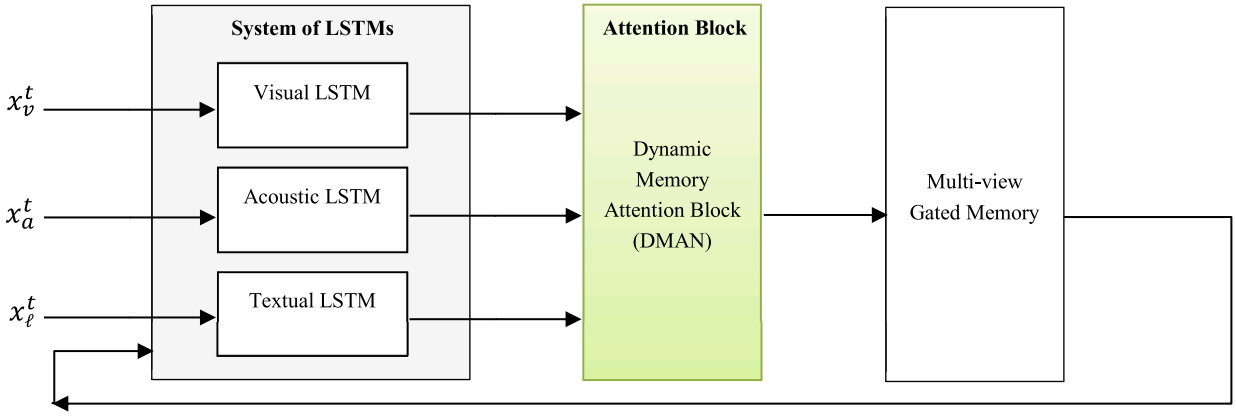

Fig. 6. Architecture of MARN.

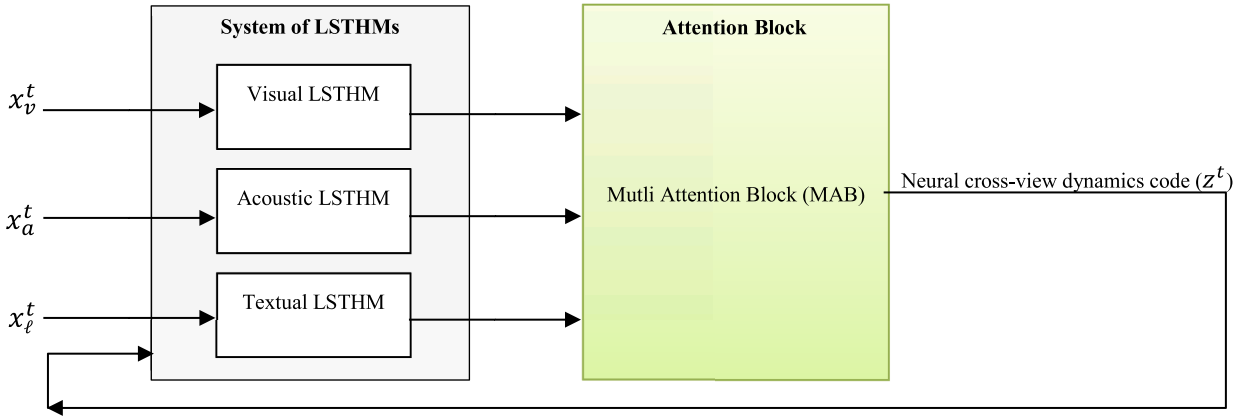

Fig. 7. Architecture of MFN.

output of this module is the attended output of the LSTHMs, which undergoes dimensionality reduction, and then passed into a deep neural network to produce the cross-view dynamics code at time (t), called z^t. The cross-view dynamics code represents all cross-modal interactions discovered at this timestep and is fed back into the intramodal LSTHMs as an additional input for the next timestep.

Finally, the cross-view dynamics code and the output of the three LSTHMs at the last timestamp of the input sequence are concatenated together to construct the multimodal sequence representation that can be used to produce the output sentiment. MARN is so powerful because it can model asynchronous cross-view dynamics.

4.3.4. *Recurrent Memory Fusion Network (RMFN)*

Similar to MARN, RMFN uses a hybrid LSTM (LSTHM) to model the view-specific dynamics [80]. For each timestep, the output of the three LSTHMs are concatenated together and passed to the attention block. The attention block in RMFN consist of multiple stages. In each stage, the most important modalities are highlighted using an Attention LSTM, then passed to a Fuse LSTM which integrates the highlighted modalities with the fusion representations from previous stages. Then the output of the final stage is fed into a SUMMARIZE module to generate a summarized cross-modal representation which represents all cross-modal interactions discovered at this timestep and is fed back into the intramodal LSTHMs as an additional input for the next timestep [80]. Finally, the last summarized cross-modal representation and the output of the three LSTHMs at the last timestamp of the input sequence are concatenated together to construct the multimodal sequence representation that can be used to produce the output sentiment.

The experiments conducted by the authors reveal that the multiple stages coordinate to capture both synchronous and asynchronous multimodal interactions.

4.3.5. *Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation*

Shenoy et al. [89] assume that the sentiment governing a specific utterance depends on 4 factors – speaker state, speaker intent, the preceding and future emotions, and the context of the conversation. The speaker intent is particularly difficult to model due to its dependency of prior knowledge about the speaker, yet modelling the other three factors separately in an interrelated manner was theorized to produce meaningful results if managed to be captured effectively. The authors attempt to simulate the setting in which an utterance is said, and use the actual utterance at that point to be able to gain better insights regarding the sentiment of that utterance. The model uses information from all modalities learning multiple state vectors (representing speaker state) for a given utterance, followed by a pairwise attention mechanism, attempting to better capture the relationship between all pairs of the available modalities. In particular, the model uses two gated recurrent units (GRU) for each modality for modelling the speaker's state and emotion. Along with these GRU's, the model also uses an interconnected context

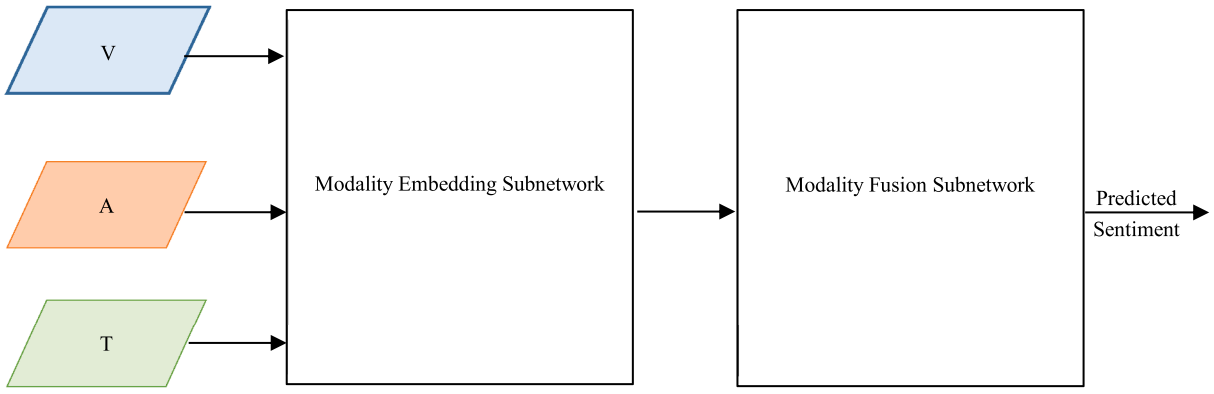

Fig. 8. Architecture of utterance-level non-temporal fusion based architecture.

network, consisting of the same number of GRU's as the number of available modalities, to model a different learned context representation for each modality. The incoming utterance representations and the historical GRU outputs are used at every timestamp to be able to predict the sentiment for that timestamp [89].

4.4. Utterance-Level Non-temporal Fusion

This architecture relies on collapsing the time dimension from inputs. Unlike architecture 3 where the model works on every time step, in this architecture, the model work with the whole utterance. We define three matrices T, A, V which are formed from the concatenation of the language features, acoustic features and visual features respectively of each utterance. Most of the models in this architecture mainly consists of two main components as shown in Fig. 8.

- Modality Embedding Subnetwork

This subnetwork is responsible for modelling the view-specific dynamics. The output from this subnetwork are acoustic, visual and textual embeddings of the utterance.

- Modality Fusion Subnetwork

The fusion layer is used to combine the acoustic, visual and textual embeddings into one compact representation to model the cross-view dynamics of the whole utterance.

4.4.1. Tensor Fusion Network (TFN)

Zadeh et al. [9] explicitly model view-specific dynamics of the textual modality with an LSTM followed by a fully connected deep network. While the view-specific dynamics of the visual and acoustic modalities are both modelled by mean pooling followed by a fully connected deep network.

Then they perform fusion by creating a multi-dimensional tensor that captures unimodal, bimodal and trimodal interactions across the three modalities. It is mathematically equivalent to the outer product between the visual embeddings, acoustic embeddings and the textual embeddings. It is a 3D cube of all possible combination of unimodal embeddings as shown in Fig. 9. So each utterance can be represented by a multimodal tensor which is passed to a fully connected deep neural network called Sentiment Inference Subnetwork to produce a vector representation which can be used to predict the sentiment (see Fig. 10).

The main disadvantage is that the produced tensor is very high dimensional and its dimensions increase exponentially with the number of modalities. Consequently the number of learnable parameters in the weight tensor in the Sentiment Inference Subnetwork will also increase

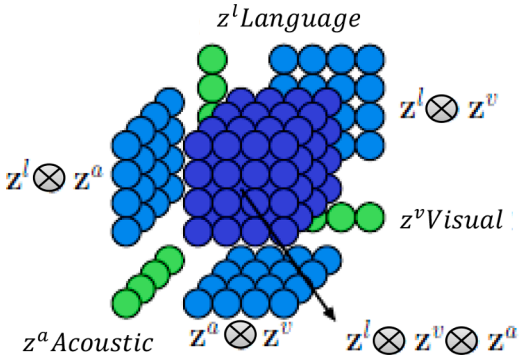

Fig. 9. Tensor fusion with three types of subtensors: unimodal, bimodal and trimodal [9].

exponentially. This introduces exponential computational increase in cost and memory and exposes the model to risks of overfitting [81] [82]. Despite the effectiveness this type of methods have achieved, they give little consideration to acknowledging the variations across different portions of a feature vector which may contain disparate aspects of information and thus fail to render the fusion procedure more specialized [83].

4.4.2. Temporal Tensor Fusion Network (T2FN)

Proceeding from the fact that clean data exhibits correlations across time and across modalities and produce low rank tensors, while noisy data breaks these natural correlations and leads to higher rank tensors, *Paul et al.* [82] propose a model called the Temporal Tensor Fusion Network (T2FN) which builds tensor representation from multimodal data but uses the method of low-rank tensor approximation to implement more efficient tensors that can represent the true correlations and latent structures in multimodal data more accurately, thus eliminating imperfection in the input. The architecture of T2FN is very similar to TFN. The main difference is that T2FN models view-specific dynamics of the three modalities with three LSTMs not followed by a deep network. And while TFN uses a single outer product between the three embeddings to obtain a multimodal tensor of rank one, T2FN performs outer products between the individual representations through every time step in each utterance to obtain a number of tensors equal to the number of time steps in the utterance. Then these tensors are summed to obtain a final tensor of high rank upper bounded by the number of time steps in each utterance. This final tensor can be used by the fully connected layer to predict the sentiment. The main advantage of these model is that the tensor rank minimization acts as a simple regularizer for training in the presence of noisy data. It is also computationally efficient and has fewer

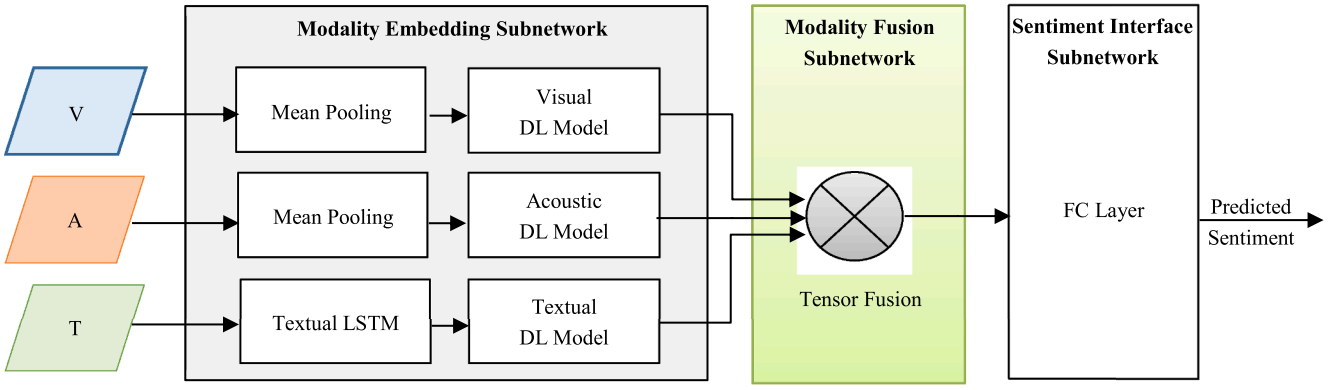

Fig. 10. Architecture of TFN.

parameter compared to previous tensor-based methods.

4.4.3. *Low-rank Multimodal Fusion (LMF)*

Similar to T2FN, Liu et al. [81] propose low-rank weight tensors to make multimodal fusion efficient without compromising on performance. The architecture of LMF is very similar to that of TFN [9]. LMF decreases the number of parameters as well as the computation complexity.

4.4.4. *Modality based Redundancy Reduction Fusion (MRRF)*

Inspired by how TFN [9] fuses multimodal by outer product tensor of input modalities, and how LMF [81] reduce the number of elements in the resulting tensor through low rank factorization, Barezi et al. [84] introduce the MRRF which builds on the above two models. The main difference is that the factorization used in LMF utilizes a single compression rate across all modalities, while MRRF use Tuckers tensor decomposition which gives different compression rates for each modality. This allows the model to adapt to variations in the amount of useful information between modalities. Modality-specific factors are chosen by maximizing performance on a validation set. Applying a modality-based factorization method is useful in removing redundant information that is duplicated across modalities and results in fewer parameters with minimal information loss leading to a less complicated model and reducing overfitting [84].

4.4.5. *Hierarchical Feature Fusion Network (HFFN)*

Again inspired by how TFN [9] fuses multimodal features by outer product tensor of input modalities, Mai et al. [83] tries to improve the efficiency and avoid the problem of high dimensional tensors being created by introducing a new model called HFFN that has proven empirically to achieve significant drop in computational complexity compared to other tensor based methods. The model consists of three main stages: 'divide', 'conquer' and 'combine' (see Fig. 11). In the 'divide' stage, the feature vectors of the three modalities of the target utterance are aligned to form multimodality embedding, then this multimodality embedding is divided into multiple local chunks using a sliding window to explore inter-modality dynamics locally. In the

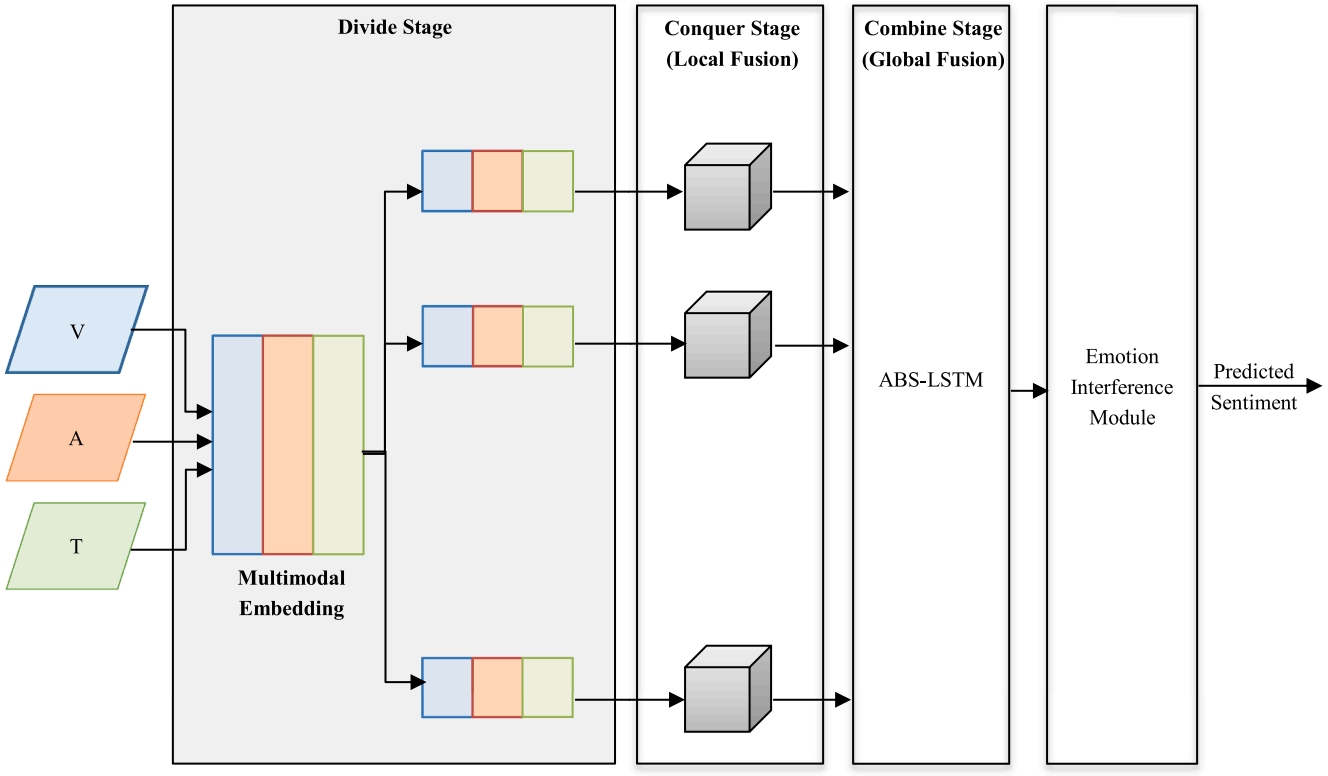

Fig. 11. Architecture of HFFN.

Conquer stage, the local chunks are passed into Local Fusion Module (LFM) where the outer product is applied for fusing features within each local chunk to model the inter-modality dynamics. This can considerably reduce the computational complexity by dividing holistic tensor of TFN into multiple local ones. In the Combine stage, global interactions are modelled by exploring interconnections and context-dependency across local fused tensors. However, the limited and fixed size of sliding window may lead to division of the complete process of expressing the sentiment into different local portions, that's why bidirectional flow of information between local tensors is warranted to compensate for this problem. This is what urged the authors to design an RNN variant, called Attentive Bi-directional Skip-connected LSTM (ABS-LSTM); it is bidirectional and supported with two levels of attention mechanism: Regional Interdependence Attention and Global Interaction Attention. ABS-LSTM can transmit information and learn cross-modal interactions more effectively. Finally the output of the global fusion module is passed to Emotion Inference Module (EIM) to predict the sentiment of the target utterance.

4.4.6. *Multimodal Uni-utterance Self Attention (MMUUSA)*

Ghosaly et al. [85] explicitly models the view-specific dynamics of the three modalities by a bidirectional GRU followed by a fully connected deep network to generate visual, acoustic and textual embeddings (modality embedding subnetwork) as shown in Fig. 12. The three embeddings are concatenated together to form the information matrix of the utterance. And in order to model the cross-modal interactions, self-attention is applied to the information matrix to produce the attention matrix. Finally, the information matrix and attention matrix are concatenated and passed to the output layer for prediction of the sentiment of each utterance.

4.4.7. *Multimodal Uni-utterance Bi-Attention (MMUUBA)*

In the same paper [85], Ghosaly et al. try modelling the cross-view dynamics using bi-modal attention. Pairwise attentions are computed across all possible combinations of modality embeddings, i.e., linguistic–visual, linguistic–acoustic, and visual–acoustic. Finally, individual modality embeddings and bimodal attention pairs are concatenated to create the multimodal representation that can be used for the prediction of the sentiment of each utterance as shown in Fig. 13.

4.4.8. *Bimodal Information-augmented Multi-Head Attention (BIMHA)*

BIMHA [86] consists of four layers. The first layer models the view specific dynamics within the single modality. The second layer models the cross-view dynamics. Wu et al. [86] adopted tensor fusion based approach, which calculates the second order Cartesian product from the embeddings of pairwise modalities to obtain the interaction information. After that, in order to adapt to the attention calculation module, the extracted unimodal features are fed into to two fully connected layers to uniformly convert the feature dimension to p. Different from the first fully connected layer which is private for each modality, the second fully connected layer is a shared layer to reduce parameters. The third layer extracts the bimodal interaction information. In this layer, the multi-head attention mechanism is applied to conduct bimodal interaction and calculate the bimodal attention to obtain the features assigned attention weights. Finally, individual modality embeddings and bimodal attention pairs are concatenated to create the multimodal representation that can be used for the prediction of the sentiment of each utterance.

4.4.9. *Sentimental Words Aware Fusion Network for Multimodal Sentiment Analysis (SWAFN)*

Very similar to the architecture of MMUUBA, Chen and Li [87] explicitly model the view-specific dynamics of the three modalities by LSTM to generate visual, acoustic and textual embeddings (modality embedding subnetwork). Then they use the coattention mechanism introduced by Xiong et al. [88] to learn the co-dependent representation between language modality and other modalities (i.e. vision or acoustic) separately by capturing attention contexts of each modality. This kind of bimodal fusion between language modality and other modalities is called the shallow fusion part of the network, as the trimodal fusion and the knowledge existing in the language modality are not well captured so far [87].

In order to capture the knowledge existing in the language modality, the authors concatenate the two kinds of bimodal fusion representation and the language embedding, and input the result to an LSTM layer to aggregate. Moreover, the authors believe that the sentimental words information that exist in the language modality can also be incorporated into a fusion model to learn richer multimodal representation. This is what urged them to design a sentimental words prediction task as an auxiliary task to guide the aggregation of the shallow fusion of multiple modality features and obtain the final sentimental words aware deep fusion representation. This part of the network is called the aggregation part and this part is mainly the main difference between MMUUBA and SWAFN. Finally, the final representation is input to a fully-connected layer and a prediction layer to get the sentiment prediction [87]. Please refer to the original paper [87] for detailed explanation of the aggregation part.

4.5. *Word Level Fusion*

In this architecture, every word in a sequence is fused with the accompanying nonverbal (acoustic and visual) features to learn variation vectors that either (1) disambiguate or (2) emphasize the existing word representations for multimodal prediction tasks [78].

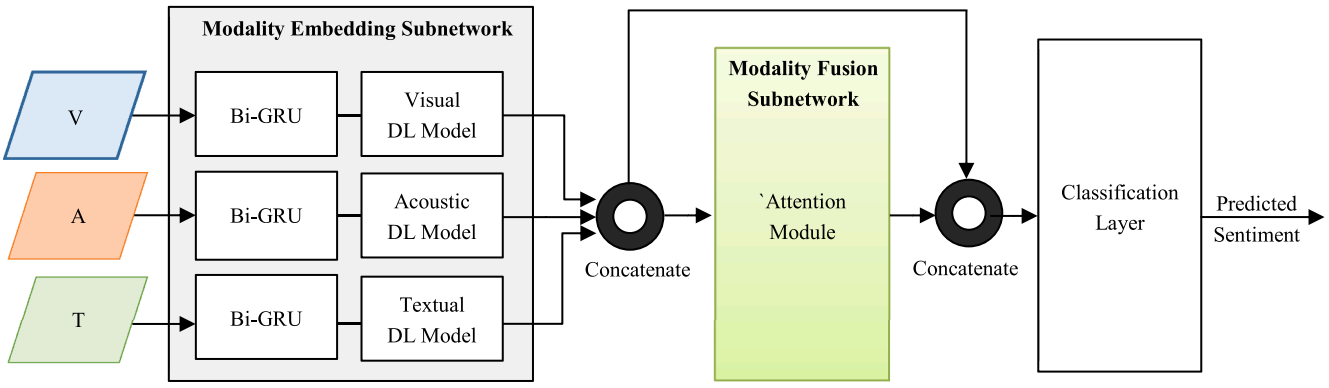

Fig. 12. Architecture of MMUUSA.

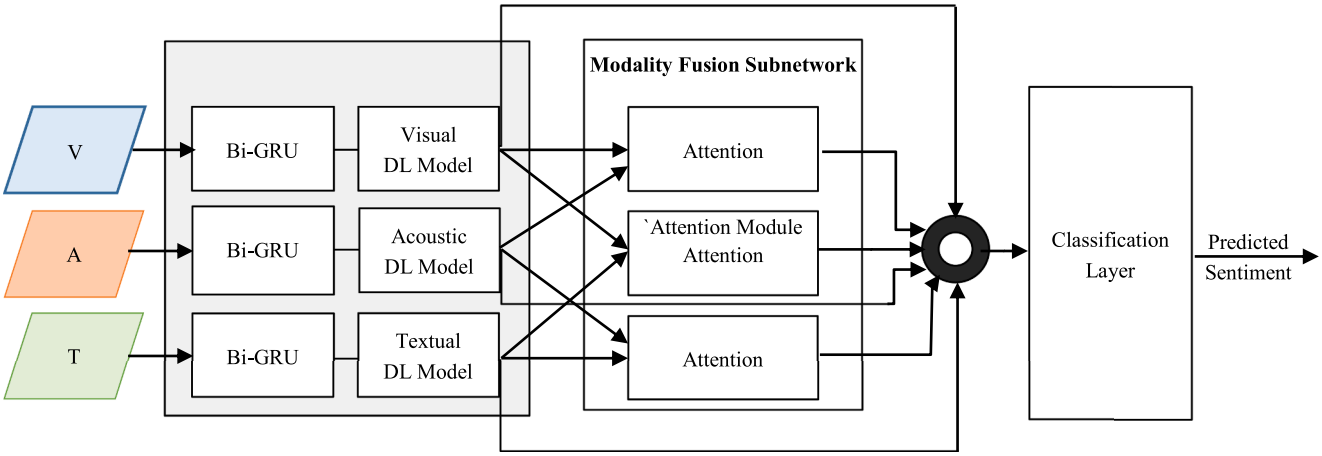

Fig. 13. Architecture of MMUUBA.

4.5.1. RAVEN

Wang et al. [78] assume that the exact sentiment behind an uttered word can always be derived from the embedding of the uttered words combined with a shift in the embedding space introduced by the accompanying nonverbal (acoustic and visual) features. Proceeding from this assumption, Wang et al. [78] use the visual and acoustic modalities accompanying an uttered word to learn variation vectors that can either disambiguate or emphasize the existing word representations for multimodal prediction tasks.

In RAVEN, each word in an utterance is accompanied by two sequences from the visual and acoustic modalities. Our model consists of three major components: (1) Nonverbal Subnetwork where the visual and acoustic features are passed into visual and acoustic LSTMs respectively to model the view-specific dynamics and compute the nonverbal (visual and acoustic) embeddings. (2) Gated Modality mixing Network takes as input the original word embedding as well as the visual and acoustic embedding, and uses an attention gating mechanism to yield the nonverbal shift vector which characterizes how far and in which direction has the meaning of the word changed due to nonverbal context. (3) Multimodal Shifting computes the multimodal-shifted word representation by integrating the nonverbal shift vector to the original word embedding.

By applying the same method for every word in a sequence, the original sequence triplet (language, visual and acoustic) is transformed into one sequence of multimodal-shifted representations (E) which corresponds to a shifted version of the original sequence of word representations fused with information from its accompanying nonverbal contexts. This sequence of multimodal-shifted word representations is then used in the high-level hierarchy to predict sentiments or emotions expressed in the utterance.

4.6. Multi-Modal Multi-Utterance Fusion

All the previous approaches consider each utterance as an independent entity and, ignore the relationship and dependencies between other utterances in the video. Utterance-level sentiment analysis and traditional fusion techniques cannot extract context from multiple utterances. But practically, utterances in the same video maintain a sequence and can be highly correlated. Thus, identifying relevant and important information from the pool of utterances is necessary in order to make a model more robust and accurate [85, 90, 91]. Proceeding from this claim, some researches recently started to benefit from the contextual information of other utterances in order to classify an utterance in video. However, every modality and utterance may not have the same importance in the sentiment and emotion classification, therefore the architecture mainly consists of two main modules whose order may vary from one model to another:

- Context Extraction Module

This module is used to model the contextual relationship among the neighbouring utterances in the video. Also, all neighboring utterances are not equally important in the sentiment classification of the target utterance. Thus it is necessary to highlight which utterances of the relevant contextual utterances are more important to predict the sentiment of the target utterance. In most architectures this module is usually a bidirectional recurrent neural network based module.

- Attention-Based Module for Modality Fusion

This module is responsible for fusing the three modalities (text, audio and video) and prioritizing only the important ones.

In the following section, we define three matrices T_v, V_v, A_v which represent the multi-modal information (i.e. text, visual & acoustic) for a sequence of utterances in a video.

4.6.1. Bidirectional Contextual LSTM (Bc-LSTM)

Poria et al. [90] propose an LSTM-based network that takes as input the sequence of utterances in a video. Initially, the unimodal features are extracted from each utterance separately without considering the contextual dependency between the utterances. For each video, a matrix is constructed from the unimodal features of all utterances in the video, then this matrix is used as input to a contextual LSTM cell such that each utterance can get contextual information from the neighboring utterances in the video. The output of each LSTM cell is passed into a dense layer followed by a softmax layer. The dense layer activations serve as the output features (see Fig. 14).

The authors [90] also consider several variants of the contextual LSTM architecture in their experiments. First they consider the simple LSTM (sc-LSTM) where the contextual LSTM architecture consists of unidirectional LSTM cells. They also consider the hidden LSTM (h-LSTM) where the dense layer after the LSTM cell is omitted. Furthermore they consider the Bi-directional LSTMs (bc-LSTM); this variant gives the best performance since an utterance can get information from utterances occurring before and after itself in the video.

4.6.2. Multi-Utterance - Self Attention (MU-SA)

Ghosaly et al. [85] extract the context between the neighboring utterances at one level using bidirectional recurrent neural networks based models. The proposed framework takes multi-modal information (i.e. text, visual & acoustic) for a sequence of utterances and feeds it into three separate bi-directional Gated Recurrent Unit (GRU) [92] (one for

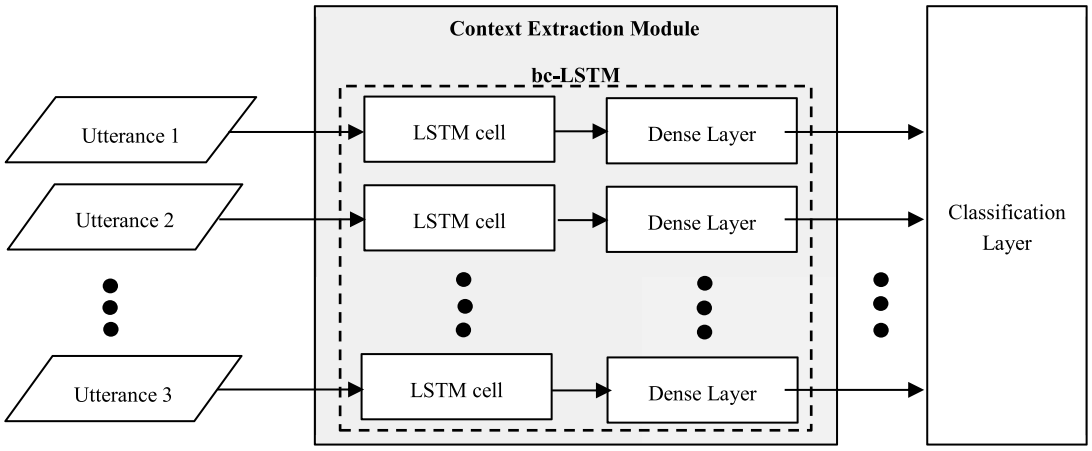

Fig. 14. Architecture of bc-LSTM.

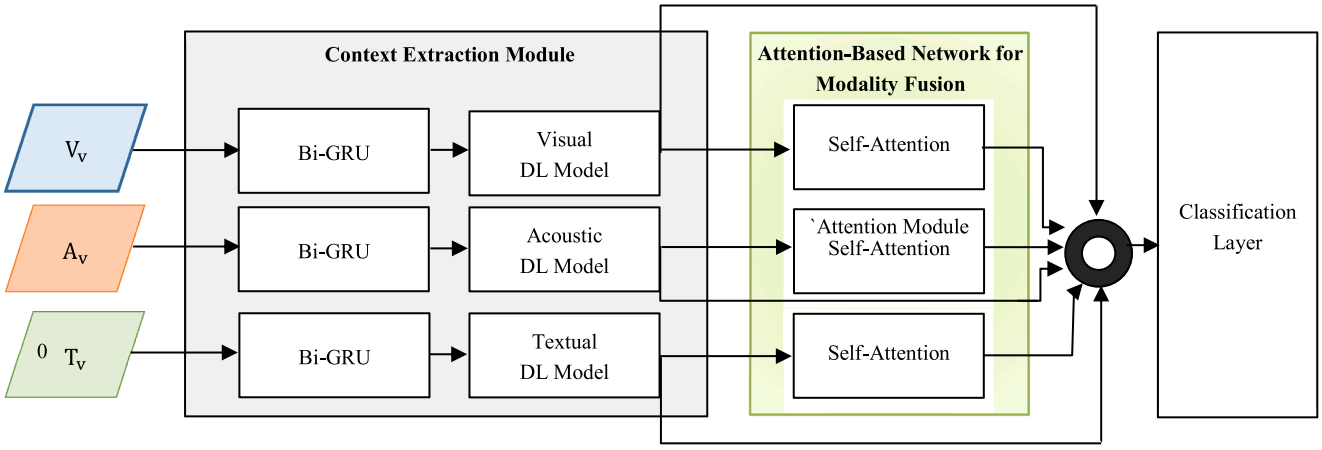

Fig. 15. Architecture of MUSA.

each modality), followed by three fully-connected dense layers (one for each modality) resulting in three matrices that contain modality specific contextual information between the utterances. This is the only level of context extraction and is called unimodal context extraction.

Next self-attention is applied on the utterances of each modality separately, and used for classification (see Fig. 15). Specifically, for the three modalities, three separate attention blocks are required, where each block takes multi-utterance information of a single modality and computes the self-attention matrix. Finally the attention matrices, along with output of the dense layers are concatenated and passed to the output layer for classification.

4.6.3. *Multi-Modal Multi-utterance Bi-Attention (MMMUBA)*

Ghosaly et al. [85] replicate the same context architecture module of

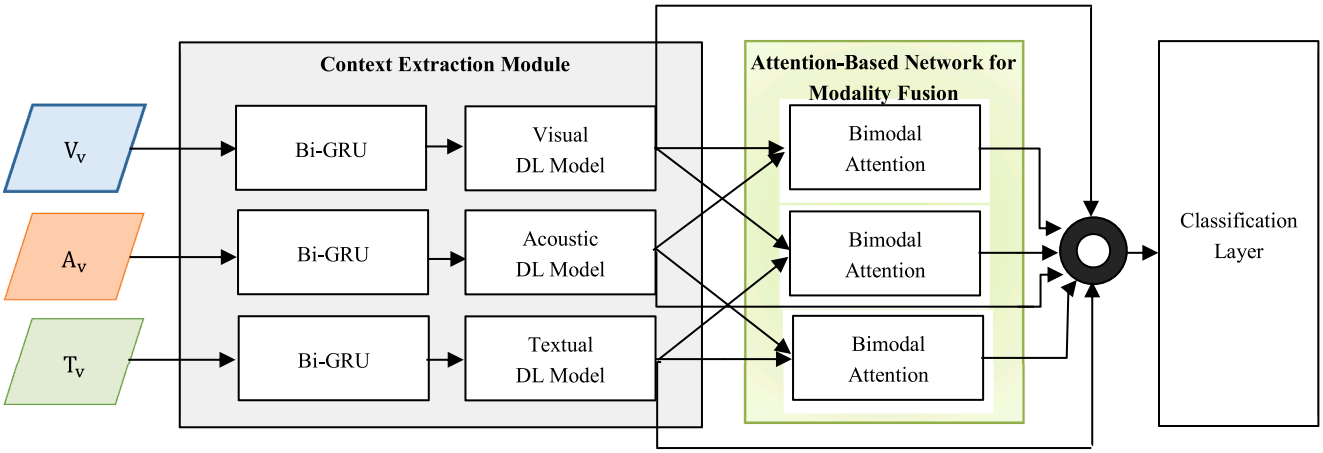

Fig. 16. Architecture of MMMUBA.

MU-SA, however in this model, multimodal attention is applied on the outputs of the dense layers in order to learn the joint-association between the multiple modalities & utterances, and to emphasize on the contributing features by putting more attention to these. In particular, bi-modal attention framework is employed, where an attention function is applied on the representations of pairwise modalities i.e. visual-text, text-acoustic and acoustic-visual to obtain bimodal representations. Finally the bimodal representations along with individual modalities are concatenated using residual skip attention-based networks [93] and passed to a softmax layer for classification as shown in Fig. 16.

4.6.4. Multi-Modal Multi-utterance Bi-Attention 2 (MMMUBA II)

Huddar et al. [94] modified MMMUBA by adding two more levels of context extraction between the neighboring utterances using bidirectional recurrent neural networks based models. The introduced model is almost identical to MMMUBA, that's why we called it MMMUBA II although the authors haven't named their proposed model explicitly. The only difference between the two models is that before the final classification, the bimodal representations are fed into a bidirectional recurrent neural network-based module in order to extract a bimodal contextual feature vector. This is the second level of context extraction and is called bimodal context extraction. The bimodal contextual feature vectors are concatenated using residual skip attention-based networks [93] to obtain a trimodal attention matrix (Trimodal Fusion), which is passed into a bidirectional LSTM to obtain trimodal contextual features. This is the third level of context extraction and is called trimodal context extraction. Finally the contextual trimodal attention matrix is fed into a softmax classifier to obtain the final classification label for the utterance.

4.6.5. Multi-Head Self-Attention Network (MHSAN)

Cao et al. [1] claim that traditional multimodal sentiment analysis methods are mainly based on RNNs which cannot utilize the correlation between each sentence well. To address this issue, they propose multimodal sentiment analysis based on the multi-head attention mechanism, considering both the context information between sentences and the contributing factors of different modalities.

The proposed framework extracts the context between the neighboring utterances at one level using multi-head attention based networks (see Fig 17). The multimodal information (i.e. text, visual & acoustic) for a sequence of utterances are fed it into three separate multi-head attention networks. Then, the context-dependent unimodal features are concatenated and fed into an attention network that can dynamically assign the contribution of multimodal information to sentiment classification.

4.6.6. Contextual Attention BiLSTM

For each utterance, the feature vectors of all the three modalities are fed into a fully-connected layer for dimensionality equalization, then they are concatenated vertically into a single vector called the multimodal feature vector of an utterance [95] (see Fig. 18).

The second layer in the model is called Attention-Based Network for Multimodal Fusion (AT-Fusion). This layer takes as input the multimodal feature vector of each utterance, and outputs the attended modality features of each utterance. The third layer is the Contextual Attention LSTM (CAT-LSTM) which is used to model the contextual relationship among utterances and highlight the important contextual information for classification. CAT-LSTM accepts the attended modality features (output of the second layer) of a sequence of utterances per video and outputs a new representation of those utterances based on the surrounding utterances. CAT-LSTM consists of a number of LSTM cells equal to the number of utterances in the sequence followed by an attention network to amplify the contribution of context-rich utterances. The output of each cell in the CAT-LSTM represents the new representation of each utterance and is sent into a softmax layer for sentiment classification.

4.6.7. Multi Attention Recurrent Neural Network (MA-RNN)

The architecture of MA-RNN [96] is almost identical to that of the Contextual Attention BiLSTM [95] except for that Kim et al. used Scaled Dot-Product Attention to calculate the attention score of each modality, and used multi-head attention mechanism to learn features in multiple representation subspaces at different positions. Second, an Attention-based BiGRU is used to model the contextual relationship among utterances instead of the CAT-LSTM that has been used in BiLSTM.

In particular, for each utterance, the feature vectors of all the three modalities are fed into a fully-connected layer for dimensionality equalization, then they are concatenated horizontally into a single matrix called the multimodal feature matrix of an utterance [96]. The second layer in the model is an attention layer to fuse the multimodal data of each utterance and for dimensionality reduction. This layer takes as input the multimodal feature matrix of each utterance and produces the attended modality features vector of each utterance. The scaled dot-product attention and the concept of multi-head attention introduced by Vaswani et al. [97] were applied in this layer. We refer the reader to [97] for a more detailed explanation of the model. The third layer is Attention-based BiGRU used to model the contextual relationship among utterances and highlight the important contextual information for classification. Attention-based BiGRU accepts the attended modality features (output the second layer) of a sequence of utterances per video and outputs a new representation of those utterances based on the surrounding utterances. Attention-based BiGRU consists of a number of Bi-GRU cells equal to the number of utterances in the sequence followed by an Attention network to amplify the contribution of

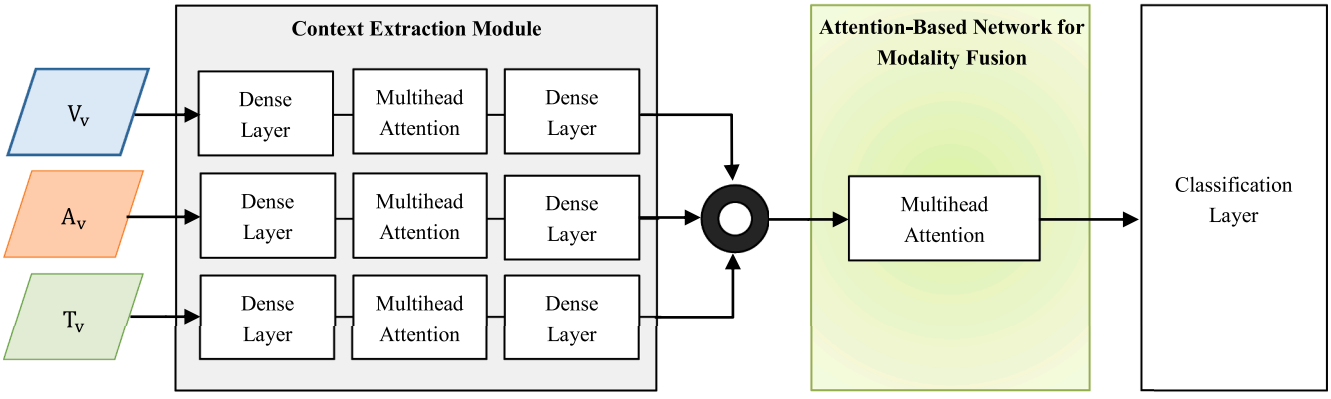

Fig. 17. Architecture of MHSAN.

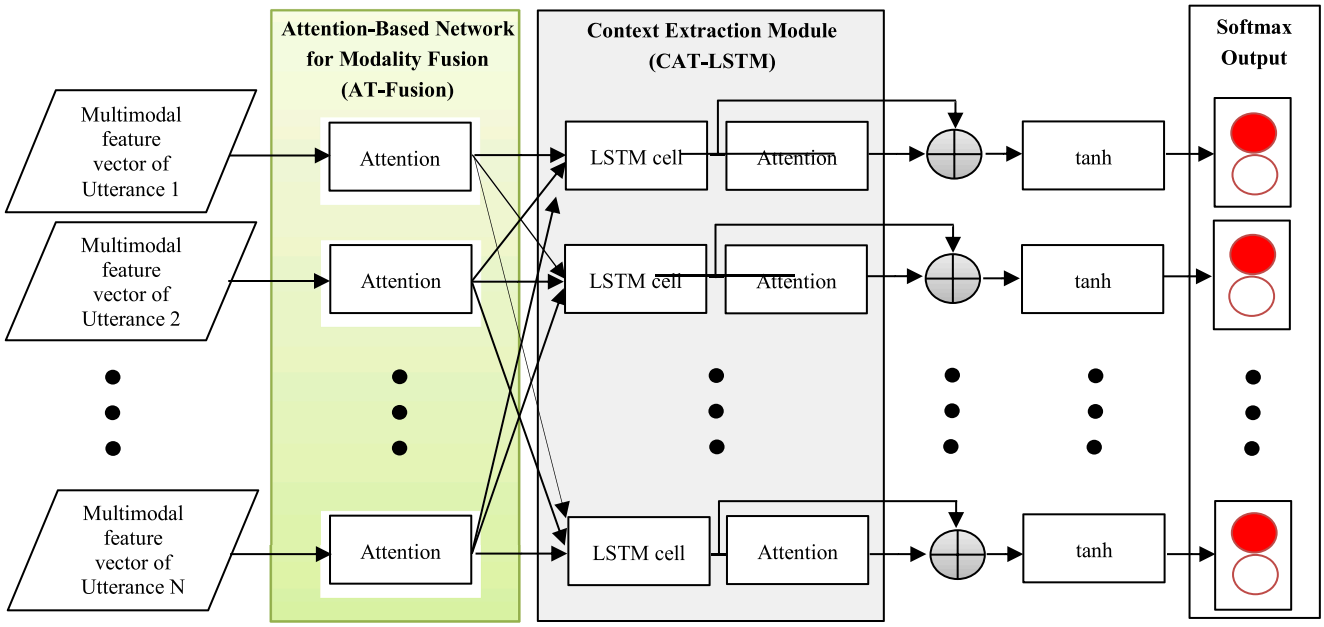

Fig. 18. Architecture of biLSTM.

context-rich utterances. The output of each cell in the Bi-GRU represents the new representation of each utterance and is sent into a softmax layer for sentiment classification [96].

4.6.8. *Multimodal sentiment analysis based on multi-head attention mechanism (MHAM)*

Xi et al. [98] proposed a model based on multi-head attention mechanism, which uses self-attention mechanism to extract the intra-modal features and the multi-head mutual attention to analyze the correlation between different modalities. And it contains a total of 6 modules derived from the multi-head attention mechanism.

4.7. *Sequence to Sequence (Seq2Seq) Models*

4.7.1. *Multimodal Cyclic Translation Network (MCTN)*

Inspired by the success of Seq2Seq models, Pham et al. [99] propose the Multimodal Cyclic Translation Network model (MCTN) to learn robust joint multimodal representations by translating between modalities. Translation from a source modality to a target modality results in an intermediate representation that captures joint information between the two modalities. MCTN extends this insight using a cyclic translation loss involving both forward translations from source to target modalities, and backward translations from the predicted target back to the source modality. This is what the authors call multimodal cyclic translations to ensure that the learned joint representations capture maximal information from both modalities The model is a hierarchical neural machine translation network with a source modality and two target modalities. The first level learns a joint representation by using back translation. This intermediate representation is translated into the second target modality without back translation. The multimodal representation is fed into RNN for final classification. The power of MCTN is that once the translation model is trained with paired multimodal data, only data from the source modality is needed at test time for final sentiment prediction which makes the model robust from perturbations or missing information in the other modalities [99].

4.7.2. *Multimodal Transformer (MulT)*

Tsai et al. [100] propose the Multimodal Transformer for Unaligned Multimodal Language Sequences (MulT). The authors extend the standard Transformer network that has been introduced by Vaswani et al. in 2017 [97] to produce a modified transformer model called the cross-modal transformer. MulT fuses multimodal time series using a feed-forward fusion process from multiple directional pairwise cross-modal transformers. The proposed cross-modal transformer has no encoder-decoder structure, however it enables one modality to receive information from another modality (i.e. tries to repeatedly reinforce a target modality with the low-level features from another source modality) by learning the attention across the two modalities' features using the cross model attention block. Each crossmodal transformer consists of several layers of crossmodal attention blocks, which can directly attend to low-level features of every pair of modalities and doesn't rely on taking intermediate-level features (removing the self-attention) which helps to preserve the low-level information for each modality. The multi-head attention block is also adopted to learn the inter-modal attention. Then, the outputs from the crossmodal transformers that share the same target modality are concatenated. Each of them is then passed through a self-attention transformer [97]. Finally, the last elements of the self-attention transformers are extracted to pass through fully-connected layers to make predictions.

4.8. *Quantum based models*

All existing multimodal sentiment analysis models that are mainly based on neural networks, model the multimodal interactions in a way that is implicit and hard-to-understand. The models suffer from low interpretability; the way the modalities interact is ambiguous and implicit for both levels of interactions. This is mainly because most models rely on neural structures to fuse multimodal data, which act like black-boxes with few numerical constraints [11]. And although the pre-mentioned models have been a success, researchers are looking for ways to understand the model, in order to know whether we can trust it and deploy it in real work, or whether it contains privacy or security issues [101]. That's why Interpretability has become an important concern for machine learning researchers and they started to develop quantum-based approaches for fusing multimodal data. As far as our knowledge three models only were introduced till 2021 that were based in quantum fusion. The first one is QMR (Zhang et al. [102] are the first to apply Quantum Theory (QT) to sentiment analysis). The second model is QMF [103] which address this limitation with inspirations from quantum theory, which contains principled methods for modeling

complicated interactions and correlations. In their quantum-inspired framework, the view-specific dynamics and the cross-view dynamics are formulated with superposition and entanglement respectively at different stages. The complex-valued neural network implementation of the framework achieves comparable results to state-of-the-art systems on both MOSI and MOSEI datasets. The third model introduced is QMN (quantum-like multimodal network) [104], which leverages the mathematical formalism of quantum theory (QT) and a long short-term memory (LSTM) network. Specifically, the QMN framework consists of a multimodal decision fusion approach inspired by quantum interference theory to capture the interactions within each utterance and a model inspired by quantum measurement theory to model the interactions between adjacent utterances. However, it is worth mentioning that QMN was tested on emotion recognition tasks, not sentiment analysis [104].

5. Performance Evaluation

We evaluate the performance of the thirty-five models in terms of both effectiveness and efficiency on CMU-MOSI and CMU-MOSEI datasets.

5.1. Effectiveness

We use five evaluation performance metrics introduced in prior work [13, 14, 73, 80, 100] to compare the effectiveness of the thirty five models on CMU-MOSI and CMU-MOSEI datasets. The five metric are:

- Binary accuracy (Acc_2)

$$\text{Sentiment} = \begin{cases} positive, \ values \geq 0 \\ negative, \ values \geq 0 \end{cases}$$

- 7-class accuracy (Acc_7)

$$Acc_7 \sim Sentiment\ score\ classification\ in Z \cap [-3, 3]$$

- F1 score (F1)

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- Mean Absolute Error (MAE)

Mean Absolute Error is a performance metric used if the problem is treated as a regression problem. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors over all instances in the test set. Each prediction error is the difference between the true value and the predicted value for the instance. Lower values denote better performance.

$$MAE\ for\ test\ set = \frac{\sum_{i=1}^{n} abs(y_i - (x_i))}{n}$$

where

y_i *is the true sentiment for test instance* x_i

(x_i) *is the predicted sentiment*

n *is the number of test examples*

- The Pearson's correlation (Corr) between the model predictions and regression ground truth.

Higher values of these metrics denote better performance for all metrics. The only exception is MAE which lower values indicate better performance.

5.2. Efficiency

The efficiency of the model is evaluated in terms of the model size (i.e., number of parameters) and the training time of learning. It is hard to compare between the models in terms of the training time since it depends heavily on the platform on which the model was trained. To compare the training time, one needs to replicate the thirty-five models on the same platform and report the training time of each, which is out of the scope of this survey. However, we will mention the results reported by *D. Gkoumas et al.* [13] who replicated the implementation of eleven of the pre-mentioned models on the same platform and reported the training time and the number of parameters of each on both MOSI and MOSEI.

6. Results and Discussion

The results mentioned in this section are either reported by the authors of each model or obtained by *D. Gkoumas et al.* [13] who replicated the implementation and reported the results of some of the pre-mentioned models for modelling human language on both MOSI and MOSEI. We also omitted the performance of very few models because it was neither reported by the authors nor by any survey paper.

For a more robust comparison between the eight architectures, we compare the effectiveness and efficiency of the models with the aid of tables and graphs. Specifically, in table 3 and table 4, we present the effectiveness results for most of the pre-mentioned models on CMU-MOSI and CMU-MOSEI datasets respectively in addition to the features extraction methods that are used in each model. Furthermore, Fig. 18 and Fig. 19 present bar charts showing the binary accuracy percentage of the referenced models on CMU-MOSI and CMU-MOSEI datasets respectively. Moreover, Fig.20 contains a bar chart showing the best binary accuracy achieved by each architecture on CMU-MOSI dataset. On the other hand, table 5 shows the efficiency results of eleven of the pre-mentioned models as reported by *D. Gkoumas et al.* [13]. In particular, table 5 presents the training time and the number of parameters on both CMU-MOSI and CMU-MOSEI datasets.

In terms of effectiveness, we observe that Majority voting model exhibits the worst performance (lowest binary accuracy and highest MAE) on both CMU-MOSI (50.2%) and CMU-MOSEI (50.2%) because modelling the cross-view dynamics is definitely more complex than a decision vote. We also observe that HMM model shows extremely low performance on CMU-MOSI. We believe that this is because HMMs were mainly introduced for the purpose of unimodal sentiment analysis and they couldn't generalize to multimodal sentiment analysis.

In general, early fusion based approaches show poor performance on CMU-MOSI. A possible reason for this may be that early fusion causes over-fitting in case of a small size training dataset like CMU-MOSI. Besides Early Fusion is not intuitively meaningful because modeling the view-specific dynamics is ignored, thus losing the context and temporal dependencies within each modality [73].

Regarding the temporal fusion based architecture, we can clearly observe that RMFN outperforms MARN and MFN on CMU-MOSI dataset, which entails that modeling the cross-modal interactions across multiple stages is beneficial.To compare multistage against independent modeling of cross-modal interactions, we pay close attention to the performance comparison between RMFN and MARN which models multiple crossmodal interactions all at once (Independent modelling of cross model interactions). RMFN shows improved performance, indicating that multistage fusion is both effective and efficient for human multimodal language modelling since the multiple stages coordinate to capture both synchronous and asynchronous multimodal interactions [80].

It has also been observed that DFG (76.9%) shows decreased

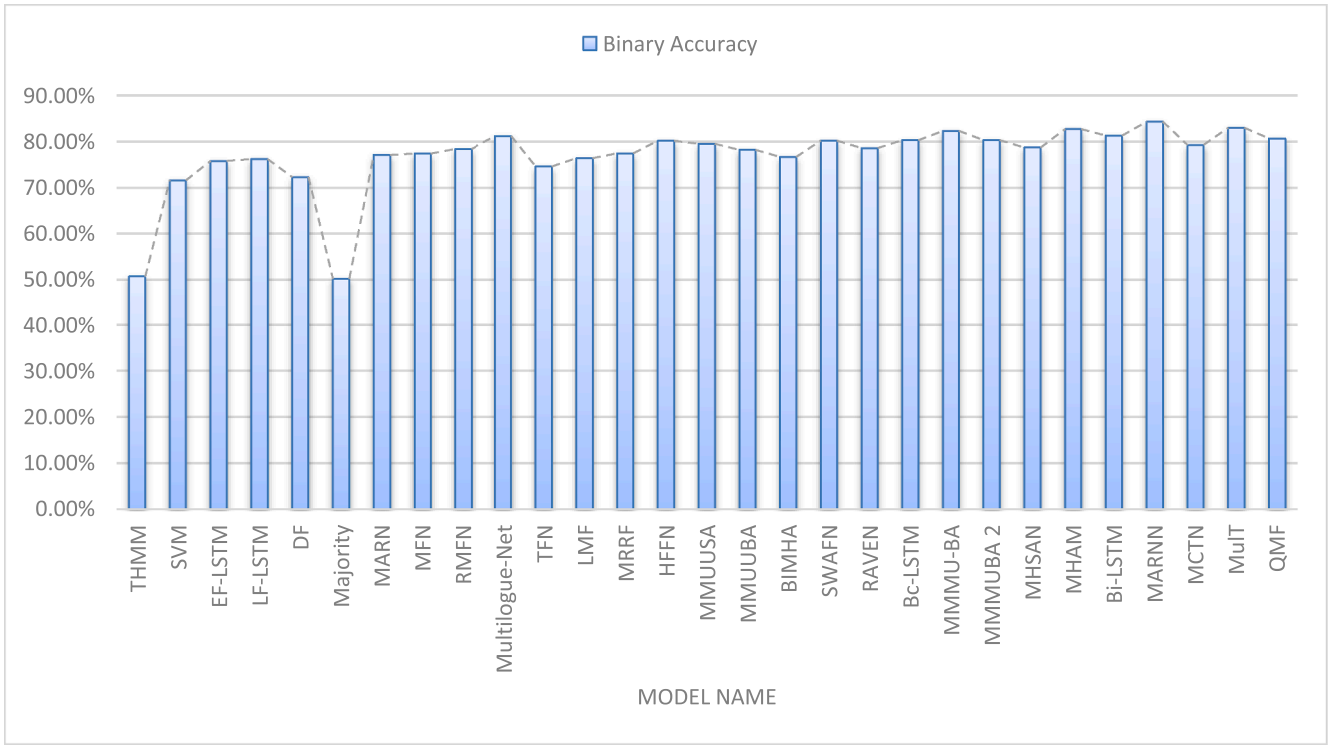

Fig. 19. Binary accuracy percentage of the referenced models on CMU-MOSI dataset.

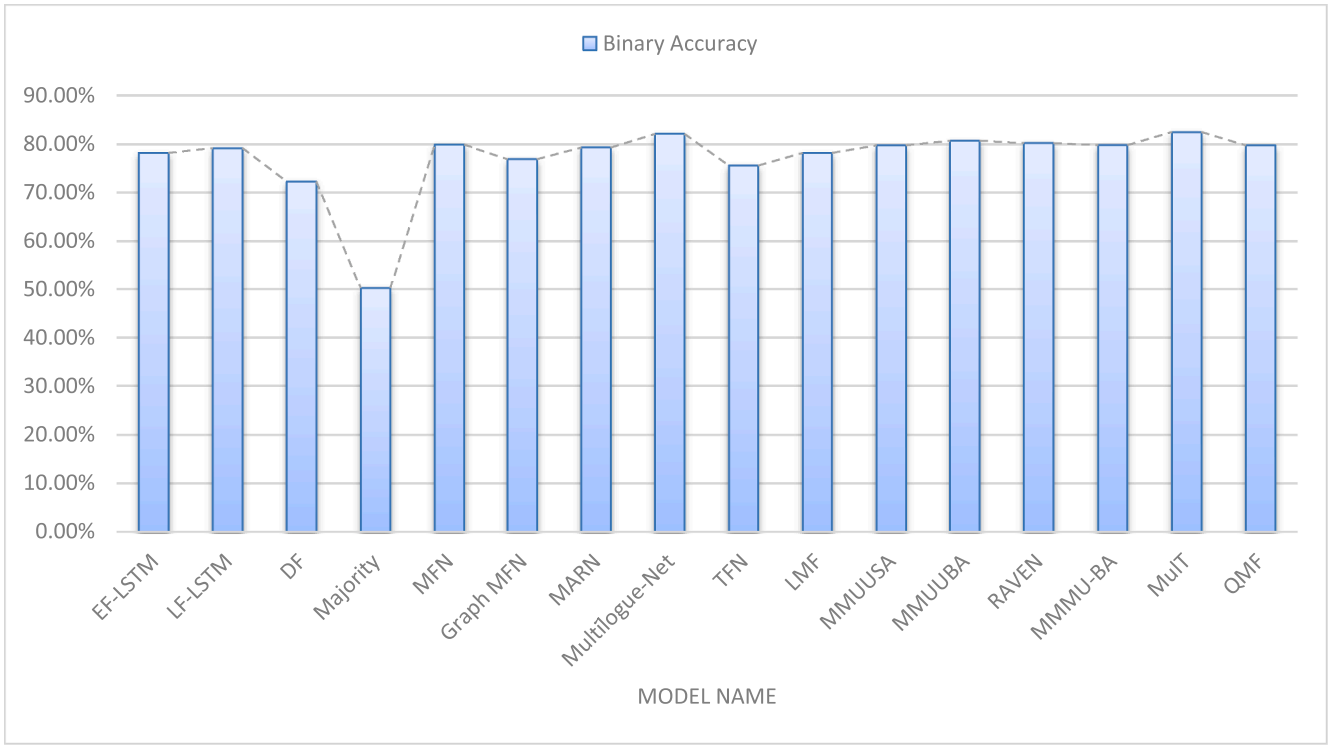

Fig. 20. Binary accuracy percentage of the referenced models on CMU-MOSEI dataset.

performance compared to MFN (79.9%) on CMU-MOSEI dataset although their architectures are very similar, which entails the crucial role of DMAN in modelling the cross-view interactions across different modalities at each timestep and that DMAN component is much more powerful than the DFG component introduced in [14].

As for Multilogue-net, it outperforms all models in the temporal fusion based architecture on both CMU-MOSI and CMU-MOSEI. We attribute this outstanding performance to two reasons. First, the authors attempt to simulate the setting in which an utterance is said by using the whole utterance information at each timestep to be able to gain better insights regarding the sentiment and emotion of that utterance [89]. Second, learning the relationship between pairs of all available

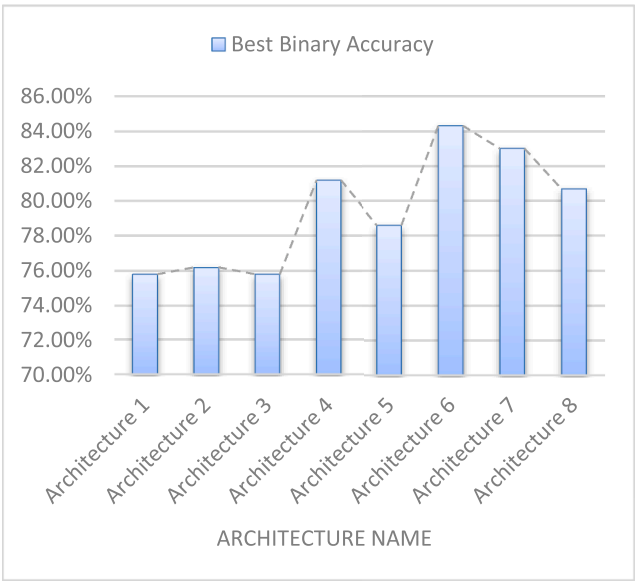

Fig. 21. Best binary accuracy percentage of each architecture on CMU-MOSI dataset.

Table 5
The number of parameters and training times in minutes for MOSI, and MOSEI of eleven models as reported by *D. Gkoumas et al.* [13].

		Efficiency			
		CMU-MOSI		CMU-MOSEI	
Ref	Model	Training Time	Number of parameters	Training Time	Number of parameters
[13]	EF-LSTM	0.43	177,329	6.59	217,457
[13]	LF-LSTM	3.14	1,155,109	54.47	5,111,485
[13]	RMFN	57.42	1,950,805	20.85	1,732,884
[13]	TFN	0.51	14,707,911	1.87	6,804,859
[13]	LMF	0.43	1,144,493	2.00	5,079,473
[13]	MARN	69.5	1,350,389	268.20	5,442,313
[13]	MulT	17.6	1,071,211	31.20	874,651
[13]	MMUU-BA	0.64	2,424,965	7.07	2,576,165
[13]	RAVEN	3.71	171,433	23.87	159,213
[13]	MFN	1.88	1,513,221	18.56	415,521
[13]	MCTN	15.64	147,100	-	-

modalities using the idea of pairwise attention introduced by Ghosal et al. [85] in 2018, where similar attributes learned by multiple modalities are emphasized and differences between the modality representations are diminished. Although using bimodal attention increases the number of parameters in the model, some researchers believe that the additional computational cost can be considered negligible in context of the increased performance [89].

Regarding the utterance non-temporal fusion based architecture, it has been observed that TFN shows very poor performance on both CMU-MOSI (74.6%) and CMU-MOSEI (75.6%) datasets. We attribute this to the very high dimensionality of the produced tensor which exposes the model to risk of overfitting especially in case of small datasets like CMU-MOSI. Despite the effectiveness this type of methods have achieved, they give little consideration to acknowledging the variations across different portions of a feature vector which may contain disparate aspects of information and thus fail to render the fusion procedure more specialized [83]. In terms of efficiency, TFN also introduces exponential computational increase in number of parameters, cost and memory (see Table 5).

On the other hand, LMF and MRRF show improved performance compared to TFN in terms of both effectiveness and efficiency, where these models apply tensor decomposition to decrease the number of parameters as well as the computation complexity, and reducing overfitting. In terms of efficiency, LMF, T2FN, and MRRF are computationally efficient and has fewer parameters compared to previous tensor-based methods.

On the flip side, HFFN (80.19%) achieves improvement on CMU-MOSI task compared with the tensor fusion approaches: TFN (74.6%), MRRF (77.46%) and LMF (76.4%) on both binary accuracy and F1 score. It is reasonable because the aforementioned methods conduct tensor fusion at holistic level and ignore modeling local interactions, while HFFN has a well-designed Local Fusion Module (LFM) [83]. In terms of efficiency also, Mai et al. [83] compare between HFFN and TFN with respect to the number of parameters and FLOPs after fusion (the FLOPs index is used to measure time complexity), and gives the same input to the two methods to make a fair comparison. The results of the conducted experiments reveal that HFFN have much less parameters and FLOPs than TFN.

As for SWAFN, it can be clearly observed that both SWAFN and HFFN achieve the best performance among the models of the 5th architecture on CMU-MOSI task on both binary accuracy and F1 score. SWAFN also achieves the best performance on mean absolute error (MAE) for regression task, and on 7-class accuracy. Due to the very limited training samples of the CMU-MOSI dataset, many baseline models may be overfitting on the training set, but SWAFN achieves better performance, indicating its better generalization ability. Several experiments conducted by Chen and Li [87] on SWAFN for all different combinations of modalities, have proven that the performance of SWAFN with auxiliary task outperform that of SWAFN without auxiliary task, which suggests that sentimental words classification auxiliary task indeed plays a remarkable role [87]. In addition, the proposed crossmodal co-attention mechanism which learns the interaction between different modalities also makes significant contribution in SFWAN. Furthermore, it can be clearly noticed that SWAFN outperforms MMUUBA in terms of the five effectiveness metrics although their architectures are very similar. The main difference between the two models is the sentimental words classification auxiliary task which again highlights the crucial role of the auxiliary task in modelling the cross-view interactions across different modalities.

As for BIMHA, it is great mentioning that it is the only model so far that has been tested on the CH-SIMS Chinese dataset. BIHMS achieves an outstanding performance in CH-SIMS in terms of all effectiveness metrics. In addition, BIMHA has a significant performance on all effectiveness metrics compared with other methods on CMU-MOSI English dataset which shows that the model has outstanding generalization ability. In terms of efficiency, BIMHA is a shallow model, and through parameter sharing, the amount of parameters and memory consumption are reduced [86].

The 6th architecture (Multimodal Multi-utterance based architecture) and MulT outperform all other architectures on CMU-MOSI. We attribute the great performance of the 6th architecture to modelling the contextual information of the neighboring utterances in order to classify the target utterance. Utterances in the same video maintain a sequence and can be highly correlated, thus, identifying relevant and important information from the pool of utterances is necessary in order to make a model more robust and accurate. However, utterance-level sentiment analysis and traditional fusion techniques cannot extract context from multiple utterances. That's why we can clearly observe that MMMU-BA model reports better accuracy (82.31%) compared to MMUUSA (79.52%) and MMUUBA (78.2%) on CMU-MOSI dataset, thus supporting our claim that multi-utterance framework (i.e. MMMU-BA) captures more information than single-utterance frameworks (i.e. MMUU-SA).

It has also been clearly observed that MMMUBA outperforms MUSA on both CMU-MOSI and CMU-MOSEI. Thus supporting our claim that bimodal attention is a better choice than self-attention mechanism; bi-

modal attention frameworks (i.e. MMMU-BA) capture more information than the self-attention frameworks [85]. Though self-attention based models achieve good results, they fail to consider the information between modalities. Indeed, extracting the important information of single modality prior to fusion, ignores the consistency and complementarity of bimodal interaction and has influences on the final decision [86]. In other words, one modality can provide additional information for the other modality, and the fusion features of any two contribute differently to the final sentiment decision. For example, it can be inferred that the person is happy when he speaks loudly with a pleasant smile. But if the spoken content expresses dissatisfaction, combining the voice and text we can judge that the person may be angry. Therefore, it is essential to find a method to better weigh the information provided by the interaction of two-two modalities in order to make the computer accurately recognize the human sentiment [86].

Moreover, it has been observed that MMMUBA outperforms MMMUBA II on CMU-MOSI dataset, which entails that adding extra levels of context extraction decreases the effectiveness of the model. We attribute such decrease in performance to overfitting. Indeed, some approaches are more prone to overfitting than others. In terms of efficiency, MMMUBA is computationally more efficient and has fewer parameters compared to MMMUBA II. Adding more context extraction levels increase the number of parameters as well as the computation complexity, and increase overfitting.

On the flip side, we attribute the poor performance of MHSAN (78.7%) to not using bidirectional recurrent neural networks in the context extraction module. It has been observed though the eight models discussed in the 6th architecture that the bidirectional recurrent neural networks are most effective for modeling the contextual information among neighboring utterances to capture the long term dependencies.

MARNN [96] shows the best performance in the 6th architecture and the best performance of the thirty-five models on CMU-MOSI (84.31%). MARNN also out-performs Bi-LSTM (82.31%) although their architectures are very similar. The main difference between the two models is the type of attention applied in each model. This entails that the scaled dot-product attention and the concept of multi-head attention [97] used in MARNN are most effective to be used in multimodal sentiment analysis task. Moreover, we noticed that MulT also achieves great performance on both CMU-MOSI (83%) and CMU-MOSEI (82.5%). This observation supports our claim regarding the scaled dot-product attention and the concept of multi-head attention.

As for the quantum based architecture, QMF obtains comparable performance to state-of-the-art models for both CMU-MOSI and CMU-MOSEI datasets and also gives an interpretation approach to facilitate the understanding of multimodal interactions from both quantum and classical perspectives. On the other hand, QMN was used for emotion recognition tasks, not sentiment analysis as mentioned earlier. Extensive experiments are conducted on two widely used datasets: the MELD and IEMOCAP datasets and the experimental results shows that QMN significantly outperforms a wide range of baselines and state-of-the-art models.

7. Conclusion and future work

This survey paper presented an overview on the recent updates in the field of multimodal sentiment analysis. The most popular datasets in the fields and the most popular feature extraction methods have been categorized and discussed. Thirty-five of the recently published and cited articles were categorized and summarized based on the architecture used in each model. The effectiveness and efficiency of the thirty-five models have been compared on two widely used datasets for multimodal sentiment analysis (CMU-MOSI and CMU-MOSEI). After carrying out an intensive analysis of the results, three conclusions have been concluded. (1) The most powerful architecture in multimodal sentiment analysis task is the Multi-Modal Multi-Utterance architecture, which is characterized by exploiting contextual information from the neighboring utterances in a video to classify the target utterance. This architecture mainly consists of two modules whose order may vary from one model to another. The first module is the Context Extraction Module that is used to model the contextual relationship among the neighbouring utterances in the video and highlight which utterances of the relevant contextual utterances are more important to predict the sentiment of the target one. In most recent models, this module is usually a bidirectional recurrent neural network based module. The second module is the Attention-Based Module that is responsible for fusing the three modalities (text, audio and video) and prioritizing only the important ones. (2) Moreover, the results obtained entailed that bimodal attention frameworks are more robust in modelling the cross-view dynamics than self-attention frameworks. Pairwise attention has proved to be incredibly effective yielding state-of-the-art performance on multimodal data with just simple representations for each modality. Although using bimodal attention increases the number of parameters in the model, some researchers believe that the additional computational cost can be considered negligible in context of the increased performance. (3) Furthermore, it has been concluded that using the scaled dot-product attention and the concept of multi-head attention are most effective for multimodal sentiment analysis task. To our knowledge, these are novel findings. We expect that the findings would provide helpful insights to the development of more effective and efficient models. In the future, we are going to focus on developing new models that are more powerful in analyzing human sentiments. We will also focus on making these models language independent in order to be able to generalize to any language during the prediction tasks. Furthermore, we intend to introduce new Arabic datasets in the field.

Credit Author Statement

Dr. Ahmed H. Yousef: Conception and design of study; Revising the manuscript critically for important intellectual content; Approval of the version of the manuscript to be published; **Eng. Sarah A. Abdu:** Acquisition of data; Analysis and/or interpretation of data; Drafting the manuscrip; Dr. Ashraf Salem: Revising the manuscript critically for important intellectual content; **Dr. Ashraf Salem:** Approval of the version of the manuscript to be published.

Declaration of Competing Interest

The authors whose names are listed immediately below certify that they have NO affi liations with or involvement in any organization or entity with any fi nancial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-fi nancial interest (such as personal or professional relationships, affi liations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

[1] R. Cao, C. Ye, H. Zhou, Multimodel Sentiment Analysis with Self-attention, Advances in Intelligent Systems and Computing, 2021, pp. 16–26.
[2] K. Sidney, J.K. D'mello, A Review and Meta-Analysis of Multimodal Affect Detection Systems, ACM, 2015.
[3] E.C. Soujanya Poria, Rajiv Bajpai, Amir Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion, 2017.
[4] S.K. D'mello, J. Kory, A Review and Meta-Analysis of Multimodal Affect Detection Systems, ACM Comput. Surv. 47 (3) (2015) 43, p. Article.
[5] E. Cambria, Affective Computing and Sentiment Analysis, IEEE Intelligent Systems 31 (2) (2016) 102–107.
[6] Onno Kampman, E.J.B., Dario Bertero, Pascale Fung. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction. in ACL. 2018. Melbourne, Australia: Association for Computational Linguistics.
[7] S. Poria, et al., Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines, IEEE Intelligent Systems 33 (6) (2018) 17–25.
[8] Amir Zadeh, R.Z., Eli Pincus, Louis-Philippe Morency, MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. 2016.

[9] M.C Amir Zadeh, Soujanya Poria, Erik Cambria, Louis-Philippe Morency, Tensor Fusion Network for Multimodal Sentiment Analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017, pp. 1103–1114.

[10] S.S. Rajagopalan, et al., Extending Long Short-Term Memory for Multi-View Structured Learning, Springer International Publishing, Cham, 2016.

[11] T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal Machine Learning: A Survey and Taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443.

[12] W. Guo, J. Wang, S. Wang, Deep Multimodal Representation Learning: A Survey, IEEE Access 7 (2019) 63373–63394.

[13] Q.L. Dimitris Gkoumas, Christina Lioma, Yijun Yu, Dawei Song, What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis, Information Fusion, 2021.

[14] P.P.L. AmirAli Bagher Zadeh, Soujanya Poria, Erik Cambria, Louis-Philippe Morency, Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph, in ACL, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2236–2246.

[15] R.M. Louis Philippe Morency, Payal 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.

[16] Perez-Rosas, V.M., R.; and Morency, L.-P., Utterance-level multimodal sentiment analysis, in ACL(1). 2013a. p. 973–982.

[17] M. Wollmer, T.Knaup F.W., B. Schuller, C. Sun, K. Sagae, L.-P. Morency, Youtube movie reviews: Sentiment analysis in an audio-visual context, Intell. Syst. IEEE 28 (3) (2013) 46–53.

[18] S.S. Park, H. S., M. Chatterjee, K.; Sagae, Multimodal Analysis and Prediction of Persuasiveness in Online Social Multimedia, ACM Transactions on Interactive Intelligent Systems 6 (3) (2016). October.

[19] W. Yu, et al., CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. in ACL, Association for Computational Linguistics, 2020.

[20] P. Ekman, W.V. Friesen, J.C. Hager, FACS investigator's guide, A human face (2002) 96.

[21] R. Ekman, What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), Oxford University Press, USA, 1997.

[22] Y. Yacoob, L. Davis, Computing spatio-temporal representations of human faces, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 70–75.

[23] M.J. Black, Y. Yacoob, Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion, in: Proceedings of IEEE International Conference on Computer Vision, 1995.

[24] Z. Zhang, Feature-Based Facial Expression Recognition: Sensitivity Analysis and Experiments with A Multilayer Perceptron, IJPRAI 13 (1999) 893–911.

[25] A. Haro, M. Flickner, I. Essa, Detecting and tracking eyes by using their physiological properties, dynamics, and appearance, in: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, 2000. CVPR 2000 (Cat. No.PR00662).

[26] M.J. Jones, T. Poggio, Multidimensional Morphable Models: A Framework for Representing and Matching Object Classes, International Journal of Computer Vision 29 (2) (1998) 107–131.

[27] T.F. Cootes, G.J. Edwards, C. Taylor, Active Appearance Models. Pattern Analysis and Machine Intelligence, IEEE Transactions on 23 (2001) 681–685.

[28] G. Donato, et al., Classifying facial actions, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (10) (1999) 974–989.

[29] Y.-L. Tian, T. Kanade, J.F. Cohn, Recognizing Action Units for Facial Expression Analysis, IEEE transactions on pattern analysis and machine intelligence 23 (2) (2001) 97–115.

[30] M.S. Bartlett, et al., Measuring facial expressions by computer image analysis, Psychophysiology 36 (2) (1999) 253–263.

[31] B. Fasel, J. Luettin, Recognition of asymmetric facial action unit activities and intensities, in: Proceedings 15th International Conference on Pattern Recognition, ICPR-2000, 2000.

[32] M.J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (12) (1999) 1357–1362.

[33] A. Lanitis, C.J. Taylor, T.F. Cootes, Automatic face identification system using flexible appearance models, Image and Vision Computing 13 (5) (1995) 393–401.

[34] T.F. Cootes, et al., Active Shape Models-Their Training and Application, Computer Vision and Image Understanding 61 (1) (1995) 38–59.

[35] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co, 1999, pp. 187–194.

[36] H. Ohta, H. Saji, H. Nakatani, Recognition of facial expressions using muscle-based feature models, in: Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170), 1998.

[37] I. Cohen, F.Gozman N.S., M.C. Cirelo, T.S. Huang, Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data, in: IEEE Computer Society Conference, 2003.

[38] S. Kimura, M. Yachida, Facial Expression Recognition and Its Degree Estimation, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[39] R. Verma, et al., Quantification of facial expressions using high-dimensional shape transformations, Journal of Neuroscience Methods 141 (1) (2005) 61–73.

[40] L. Morency, J. Whitehill, J. Movellan, Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation, in: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, 2008.

[41] C. Davatzikos, Measuring biological shape using geometry-based shape transformations, Image and Vision Computing 19 (1) (2001) 63–74.

[42] M. Pantic, L.J.M. Rothkrantz, Expert system for automatic analysis of facial expressions, Image and Vision Computing 18 (11) (2000) 881–905.

[43] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition 36 (1) (2003) 259–275.

[44] M. de Meijer, The contribution of general features of body movement to the attribution of emotions, Journal of Nonverbal Behavior 13 (4) (1989) 247–268.

[45] A.S. Stefano Piana, Antonio Camurri, Francesca Odone, A set of Full-Body Movement Features for Emotion Recognition to Help Children affected by Autism Spectrum Condition IDGEI International Workshop (2013).

[46] Piana, S., et al., Real-time Automatic Emotion Recognition from Body Gestures. 2014.

[47] G. Caridakis, et al., Multimodal emotion recognition from expressive faces, body gestures and speech, Boston, MA: Springer US, 2007.

[48] T. Balomenos, et al., Emotion Analysis in Man-Machine Interaction Systems, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[49] Xu, C., et al., Visual Sentiment Prediction with Deep Convolutional Neural Networks. 2014.

[50] Q. You, et al., Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks, in: Proceedings of the AAAI Conference on Artificial Intelligence 29, 2015.

[51] D. Tran, et al., Learning Spatiotemporal Features with 3D Convolutional Networks, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[52] S. Poria, et al., Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016.

[53] Littlewort, G., et al., The Computer Expression Recognition Toolbox. 2011. 298-305.

[54] I.R. Murray, J.L. Arnott, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, J Acoust Soc Am 93 (2) (1993) 1097–1108.

[55] fish 0,punct]"> L. Devillers, L. Vidrascu, L. Lamel, Challenges in real-life emotion annotation and machine learning based detection, Neural Networks 18 (4) (2005) 407–422.

[56] T. Vogt, E. Andre, Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition, in: 2005 IEEE International Conference on Multimedia and Expo, 2005.

[57] R.W. Levenson, Human emotion: A functional view, The nature of emotion: Fundamental questions 1 (1994) 123–126.

[58] D. Västfjäll, M. Kleiner, Emotion in product sound design, in: Proceedings of Journées Design Sonore, 2002, pp. 1–17.

[59] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (3) (2011) 572–587.

[60] E. Navas, I. Hernaez, L. Iker, An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS, IEEE Transactions on Audio, Speech, and Language Processing 14 (4) (2006) 1117–1127.

[61] N. Anand, P. Verma, Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data, Stanford University, 2015. Technical Report.

[62] Han, K., D. Yu, and I. Tashev, Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. 2014.

[63] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, Association for Computing Machinery, 2010, pp. 1459–1462.

[64] G. Degottex, et al., COVAREP — A collaborative voice analysis repository for speech technologies, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[65] F. Eyben, M. Wöllmer, B. Schuller, OpenEAR — Introducing the munich open-source emotion and affect recognition toolkit, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009.

[66] L. Zhang, S. Wang, B. Liu, Deep Learning for Sentiment Analysis : A Survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2018) 8.

[67] J. Pennington, R. Socher, C. Manning, GloVe: Global Vectors for Word Representation, Association for Computational Linguistics, Doha, Qatar, 2014. ACL.

[68] T. Mikolov, et al., Efficient Estimation of Word Representations, Vector Space, 2013.

[69] L.E. Baum, Statistical inference for probabilistic functions of finite state markov chains, The annals of mathematical statistics 37 (6) (1966) 1554–1563.

[70] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[71] S. A.W. Quattoni, L.-P. Morency, M.; Collins, T Darrell, Hidden conditional random fields IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 1848–1852.

[72] S. Hochreiter, J Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[73] P.P.L. Amir Zadeh, Mazumder Navonil, Poria Soujanya, Cambria Erik, Morency Louis-Philippe, Memory Fusion Network for Multi-View Sequential Learning, in: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018.

[74] P.P.L. Amir Zadeh, Soujanya Poria, Prateek Vij, Erik Cambria, Louis-Philippe Morency, Multi-attention Recurrent Network for Human Communication Comprehension, in: 32nd AAAI Conference on Artificial Intelligence, AAAI, 2018, 2018.
[75] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (8) (1997) 1735–1780.
[76] S.W. Minghai Chen, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, Louis-Philippe Morency, Multimodal Sentiment Analysis withWord-Level Fusion and Reinforcement Learning, ACM, 2017.
[77] T.W.a.S. Scherer, What really matters — An information gain analysis of questions and reactions in automated PTSD screenings, in: Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, 2017, pp. 15–20.
[78] Yansen Wang, Y.S., Zhun Liu, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors, in The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19).
[79] D. B.G. Nojavanasghari, J. Koushik, T.; Baltrušaitis, L.-P. Morency, Deep multimodal fusion for persuasiveness prediction, in: In Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI, NY, USA, ACM, 2016, pp. 284–288, 2016: New York
[80] Z.L. Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, Multimodal Language Analysis with Recurrent Multistage Fusion (RMFN, ACL, Florence, Italy, 2018.
[81] Y.S. Zhun Liu, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), Melbourne, Australia, 2018, pp. 2247–2256.
[82] Z.L Paul Pu Liang, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, Louis-Philippe Morency, Learning Representations from Imperfect Time Series Data via Tensor Rank Regularization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 1569–1576. July 28 - August 2.
[83] S. Mai, H. Hu, S. Xing, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing, Divide, ACL, 2019.
[84] J. Elham, P.F. Barezi, Modality-based Factorization for Multimodal Fusion. ACL, Association for Computational Linguistics, Florence, Italy, 2019, pp. 260–269.
[85] M.S.A Deepanway Ghosaly, Dushyant Chauhany, Soujanya Poria, Asif Ekbaly and Pushpak Bhattacharyyay, Contextual Inter-modal Attention for Multi-modal Sentiment Analysis, ACL, Italy, 2017.
[86] T. Wu, et al., Video Sentiment Analysis with Bimodal Information-augmented Multi-Head Attention, CoRR, 2021 abs/2103.02362.
[87] Chen, M. and X. Li, SWAFN: Sentimental Words Aware Fusion Network for Multimodal Sentiment Analysis. 2020.
[88] C. Xiong, V. Zhong, R. Socher, Dynamic Coattention Networks For Question Answering, arXiv, 2016, 1611.01604.
[89] A. Shenoy, A. Sardana, Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis, Conversation, 2020.
[90] Poria, S., et al. Context-Dependent Sentiment Analysis in User-Generated Videos. in ACL. 2017.
[91] J.W. Ronan Collobert, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research (2011) 2493–2537.
[92] B.v.M. Kyunghyun Cho, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. ACL, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1724–1734.
[93] K. He, et al., Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[94] M.G. Huddar, S.S. Sannakki, V.S. Rajpurohit, Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM, Multimedia Tools and Applications 80 (9) (2021) 13059–13076.
[95] E.C. Soujanya Poriaa, Devamanyu Hazarikac, Navonil Mazumderd, Amir Zadehe, Louis-Philippe Morencye, Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis, in: IEEE International Conference on Data Mining, 2017.
[96] Taeyong Kim, Multi-Attention Multimodal Sentiment Analysis, in: ICMR '20: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 436–441.
[97] N.S. Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Advances in neural information processing systems, 2017, pp. 5998–6008.
[98] Xi, C., G. Lu, and J. Yan, Multimodal sentiment analysis based on multi-head attention mechanism. 2020. 34-39.
[99] P.P.L. Hai Pham, Thomas Manzini, Louis-Philippe Morency, Barnabàs Poczos, Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities, in: The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), 2019.
[100] S.B Yao-Hung Hubert Tsaiy, Paul Pu Liangy, J. Zico Kolteryz, Louis-Philippe Morencyy, Ruslan Salakhutdinovy, Multimodal Transformer for Unaligned Multimodal Language Sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, c 2019 Association for Computational Linguistics, 2019, pp. 6558–6569.
[101] Z.C. Lipton, The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57.
[102] D.S. Yazhou Zhang, Peng Zhang, Panpan Wang, Jingfei Li et al, A quantum-inspired multimodal sentiment analysis framework, Theoretical Computer Science 752 (2018) 21–40.
[103] D.G. Qiuchi Li, Christina Lioma, Massimo Melucci, Quantum-inspired multimodal fusion for video sentiment analysis, Information Fusion 65 (2021) 58–71.
[104] Y. Zhang, et al., A Quantum-Like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis, Information Fusion 62 (2020) 14–31.
[105] Denk, T. and A. Ramallo, Contextual BERT: Conditioning the Language Model Using a Global State. 2020.
[106] McFee, B., et al., librosa: Audio and Music Signal Analysis in Python. 2015. 18-24.