Winter 12-3-2021

# Multimodal Sentiment Analysis Based on Deep Learning: Recent Progress

Xudong Luo

Jie Liu

Pingping Lin

Yifan Fan

Luo, X.Y., Liu, J., Lin, P. & Fan, Y. (2021). Multimodal sentiment analysis based on deep learning: Recent progress. In *Proceedings of The International Conference on Electronic Business, Volume 21* (pp. 293-303). ICEB'21, Nanjing, China, December 3-7, 2021.



# Multimodal Sentiment Analysis Based on Deep Learning: Recent Progress

Xudong Luo [1,2,*]
Jie Liu [1,3]
Pingping Lin [1,4]
Yifan Fan [1,5]

*Corresponding author
[1] Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin, China,
[2] Professor, luoxd@mailbox.gxnu.edu.cn
[3] Student, 625650997@qq.com
[4] Student, 1041019568@qq.com
[5] Student, 1524905845@qq.com

**ABSTRACT**

Multimodal sentiment analysis is an important research topic in the field of NLP, aiming to analyze speakers' sentiment tendencies through features extracted from textual, visual, and acoustic modalities. Its main methods are based on machine learning and deep learning. Machine learning-based methods rely heavily on labeled data. But deep learning-based methods can overcome this shortcoming and capture the in-depth semantic information and modal characteristics of the data, as well as the interactive information between multimodal data. In this paper, we survey the deep learning-based methods, including fusion of text and image and fusion of text, image, audio, and video. Specifically, we discuss the main problems of these methods and the future directions. Finally, we review the work of multimodal sentiment analysis in conversation.

*Keywords*: Multilingual, sentiment analysis, pre-trained language models, BERT, GTP.

## INTRODUCTION

Sentiment Analysis (SA) is a process of analyzing, processing, inducing, and reasoning about personal texts with sentimental colors (Liu & Zhang, 2012). Hence, SA can help businesses improve their products, people make choices of products, and governments understand people's preferences. With the development of social networks, more and more users tend to use multiple media forms (i.e., texts plus images, animated pictures, emoticons, songs, and video) to express their opinions and sentiments. This is because "a picture is worth a thousand words," although it cannot express sentiment independently of the text but serves as an auxiliary part to remind the salient content in the text (Truong & Lauw, 2019). Generally, psychological research on sentiments shows that the pitch, intensity, speed of speech, and speech quality in speech play an essential role in SA (Hamsa *et al.*, 2020). So, it is reasonable to expect that SA on the fused information from multiple modalities is better than SA on texts only.

Thus, to understand better people's attitudes or views on certain events or topics, it is necessary for researchers to pay more attention to the SA of multimodal contents fused with text, image, and audio (Huddar *et al.*, 2020a, b; Harish & Sadat, 2020). Multimodal SA (MSA) uses Natural Language Processing (NLP), information fusion, machine learning, and Deep Learning (DL) methods to identify the subjective attitude of users through features extracted from multimodal data (text, image, voice, and video data) (Ji *et al.*, 2015; Abburi *et al.*, 2016). More specifically, MSA aims to identify and classify users' attitudes and opinions from all the aspects of words, visual and auditory. Its ultimate goal is to improve the accuracy of sentiment classification and to achieve the best prediction.

Recently, Deep learning (DL) played a vital role in image recognition, speech processing, and NLP (Soleymani *et al.*, 2017). This is because DL gets rid of the constraints of feature engineering and can achieve good output through training and optimization of parameters, and it has good domain adaptability. Basically, these problems are some classification problems. SA is also a problem of complex and subjective cognitive classification. So, surely DL technology can help SA (Xu *et al.*, 2019; Kumar *et al.*, 2020; Cai *et al.*, 2019), especially MSA (Zhang *et al.*, 2020a).

In this paper, we survey various DL-based methods for MSA. The three main problems in MSA are multimodal representation learning, multimodal alignment, and multimodal information fusion. The third is the core challenge in MSA because (1) the information of different modalities may not be completely aligned in time, some modal signals are dense, and some modal signals are sparse; (2) it is difficult for fusion models to utilize the complementarity between modalities; and (3) the type and intensity of noise in different modal data may differ. So in this paper, we focus on the fusion methods.

The current multimodal information fusion methods mainly include two types: feature-level fusion (Pérez-Rosas *et al.*, 2013) and decision-level fusion (Yu *et al.*, 2020; Zhang *et al.*, 2020c). Besides, there is a combination of the two, which not only

retains the advantages of the two but also overcomes their shortcomings (Harish & Sadat, 2020). Table 1 summarises the advantages and disadvantages of different fusion mechanisms.

Table 1: Multimodal information fusion classified by level.

| Mechanism | Input | Approach | Output | Advantage | Disadvantage | Classifier |
|---|---|---|---|---|---|---|
| Feature-level fusion (Pérez-Rosas *et al.*, 2013) | Monomodal eigenvector | Feature fusion unit | Multimodal eigenvector | grasp the correlation between multimodal features. | cannot merge modal features directly, need shared space | Single classifier |
| Decision-level fusion (Yu *et al.*, 2020) | Monomodal eigenvector | extract and classify the features of each modal to obtain the local decision result and then fuse local decision results | the final decision vector | simpler and more free can produce locally optimal results | Time-consuming | Multiple classifiers |
| Feature/decision level fusion strategies (Harish & Sadat, 2020) | Monomodal eigenvector | Feature layer fusion and decision layer fusion | Decision vector | grasp the correlation between multi-modal features and produce local optimal results | - | Multiple classifiers |

Source: This study.

According to the specific operation classification in the information fusion process, the information fusion methods mainly are simple oration-based fusion, attention-based fusion, and bilinear pooling-based fusion (Zhang *et al.*, 2020a). Simple oration-based fusion (Anastasopoulos *et al.*, 2019), such as cascade and weighted sum operations, has the advantage of weak parameter correlation and less impact. The attention mechanism applied to information fusion is that the feature vectors of different modalities are weighted differently to generate the final joint feature vector representation. Training of weight parameters is the key. The bilinear pooling method (Tenenbaum & Freeman, 2000) creates a joint representation space by calculating the image and text feature vectors. However, the model's dimensionality based on the bilinear pooling is very high, and the associated model can only be effectively trained when dealing with high-dimensional problems.

There exist some surveys on SA, but they are different from ours. From the aspects of Regarding linguistic, audio, and visual features, Jazyah and Hussien (2018) survey the studies of SA before 2018. Kaur and Kautish (2019) compare the studies of SA before 2019. However, both of them do not concern SA tasks on datasets of multiple modalities. Nor do our two surveys on SA published in 2020 (Lin & Luo, 2020a,b)(which focuses are on machine learning-based and SA application, respectively). In our survey (Lin *et al.*, 2020) in 2020, Section VI is entirely devoted to MSA. However, the eight studies mentioned there are different from ours in this paper, and the tables in the two papers are different. Moreover, this paper graphically compares the performances of various methods, but the previous one did not. In addition, in this paper, we summarised the latest work on MSA in 2021, but the previous does not. Soleymani *et al.* (2017) survey studies of MSA, but all of them are before 2017. Rather, our survey in this paper covers the work in 2020 and 2021, and our focus is on deep learning-based methods for handling three different MSA tasks. Although Zhang *et al.* (2020a) survey some models and training methods used for multimodal intelligence, they focus on the fusion of NLP and visual information rather than MSA. Baltrušaitis, Ahuja, and Morency (2018) investigate the state-of-art in multimodal machine learning and introduce general algorithms, but they are concerned little about MSA. Abdu, Yousef, and Salem (2021) summarised the multimodal video sentiment analysis based on deep learning methods. Zhang, Wang, and Liu (2018) provide a comprehensive survey of DLbased SA methods published before 2019, while our survey in this paper covers the methods proposed in 2019-2021. And ours focus on MSA, but theirs does not. Dang, Moreno--Garcìa, and De la Prieta (2020) review the latest research on DL-based for general SA rather than MSA, which is our focus in this paper. The survey of Prabha and Srikanth (2019) focuses on the various flavors of the deep learning methods used in different applications of sentiment analysis at the sentence level and aspect/target level. It concerns only the modality of text in SA, but ours in this paper concert multimodality in SA.

The rest of this paper is organized as follows. Firstly, we review DL-based methods for static MSA on text and image. Secondly, we brief DL-based methods for dynamic MSA on text, image, audio, and video. Thirdly, we discuss MSA in conversation. Finally, we conclude this paper.

## FUSION OF TEXT AND IMAGE

This section mainly discusses static MSA on text and image.

### Multi-Interaction Memory Network

To capture the impact of aspects on text and images and the multiple interactions associated with text and images, Xu *et al.* (2019) propose a Multi-Interaction Memory Network (MIMN) aspect-based MSA. The model can learn not only the

interactive influence between cross-modal data but also the self-influence of single-modal data. The model consists of two interactive memory networks that model text and image data to supervise given aspects of text information and visual information, respectively. They also construct a new publicly available multimodal data set from data captured from ZOL.com (China's leading IT information and business portal). Their experiments show that MIMN outperforms the state-of-art text-level SA methods. It can be seen from the experiment that the model proposed by them is better than all the comparison methods, and the best accuracy is 61.59%, and the Macro-F1 is 60.51%. In the future, it is worth doing more research on the intersection of aspect level and MSA.

**Deep Multimodal Attention Fusion**

To exploit the discriminative features and the internal correlation between visual and semantic contents in social medial data, Huang *et al.* (2019) propose a deep multimodal attention fusion model. First, it uses the semantic attention mechanism to learn the sentiment classifier of texts and uses the visual attention mechanism to learn the sentiment classifier of images. The goal of these two classifiers is to capture the sentiment-related words in a text and the sentiment-related information in an image area. Second, to use the complementary information in different modalities, they propose a fusion multimodal attention model for mining the correlation between different modal features. Finally, they use intermediate fusion and post-fusion strategies to integrate visual attention, semantic attention, and multimodal attention for sentiment prediction. On the data sets of Getty, Twitter, and Flickr, their experiments show that their model outperforms a state-of-the-art DL method for SA (Zadeh *et al.*, 2017) in terms of accuracy. And the accuracy of the data set Getty, Twitter, Flickr-w, and Flickr-m is 86.9%, 76.3%, 85.9%, and 85.9%, respectively. In the future, it is worth further exploring fine-grained MSA and studying whether their model can be extended to deal with other modal data (such as audio and video).

**VistaNet**

To rely on visual information as alignment for identifying the important sentences of a document using attention, Truong, and Lauw (2019) propose an optical attention network for MSA, namely VistaNet. They believe that images only play a supplementary role in the SA of comments because images cannot tell the whole story alone. So, they do not directly use images for feature classification but can play another role. In other words, as a visual means to divert attention to the most prominent sentence or aspect in a comment. VistaNet has a three-tier architecture that summarises the representation from words to sentences, then to image-specific document representation, and finally to the final document representation. The model is tested on restaurant reviews, and it may be extendable to other types of web documents, such as blog posts, tweets, or any documents containing images. Their experiment that their accuracy rate reaches 61.88%.

**CNN based method**

To take advantage of the combination of deep learning networks and machine learning to deal with text and vision, Kumar *et al.* (2020) propose a method for MSA on written texts and still images. It consists of the four modules as follows. (1) Discretisation: It employs Google Lens to separate the text from an image, then sends the text and the image to the text analytics and image analytics modules, respectively. (2) Text analytics: It determines the sentiment using a hybrid of a Convolutional Neural Network (CNN). (3) Image analytics: It uses a support vector machine (SVM) classifier (trained using bag-of-visual words) to predict the visual content sentiment. (4) Decision module: It validates and categorizes the output based on five fine-grained sentiment categories: highly positive, positive, neutral, negative, and highly negative. Their experiments show that their accuracy is nearly 91%, which is better than those by the text and image modules separately.

**BiLSTM based method**

To detect multimodal sarcasm in tweets consisting of texts and images in Twitter, Cai *et al.* (2019) created a new dataset for multimodal Twitter satire detection, and they propose a multimodal hierarchical fusion model for MSA. In the task of modal feature extraction, they first extract image features and attribute features and then use image features and attribute features combined with Bidirectional LSTM (BiLSTM) to extract text features. In the multimodal information fusion task, they use a weighted average method to fuse the eigenvectors of single-modal to generate multimodal eigenvectors. The weighted average process improves the characterization of each modal feature and is better than the ordinary single modal feature splicing effect. Their experiments show that image attributes can make up for the blank high-level abstract information between text features and image features. In their own data set, the accuracy rate reached 83.44%.

Huddar et al. (2021b) propose a method for multimodal contextual fusion. First, it uses a BiLSTM with an attention model to extract important contextual information among the utterances. Then it fuses two-two modalities at a time. Finally, it fuses all three modalities. Their experiments show that their method outperforms the existing methods by over 3% on datasets IEMOCAP and over 2% on dataset CMU-MOSI.

**EF-Net**

To process and analyze different aspects of distinguishing modalities that correspond to one target, Wang *et al.* (2021) propose an Attention Capsule Extraction and Multi-head Fusion network (EF-Net) for MSA. This model uses a two-way GRU and multi-head self-attention mechanism (MHA) to encode the semantic information of the text, using the ResNet-152 model and

capsule. The network processes images, and the integration of MHA and capsule network aims to capture the interaction between multimodal inputs. In addition to the target aspect, information from context and images is also used for emotional transmission. They manually annotated the two public TMSC datasets, Twitter15 and Twitter17. On the TABMSA task, EF-Net significantly outperforms the baseline model, with accuracy rates of 73.65% and 67.77% on the Twitter15 and Twitter17 datasets, respectively.

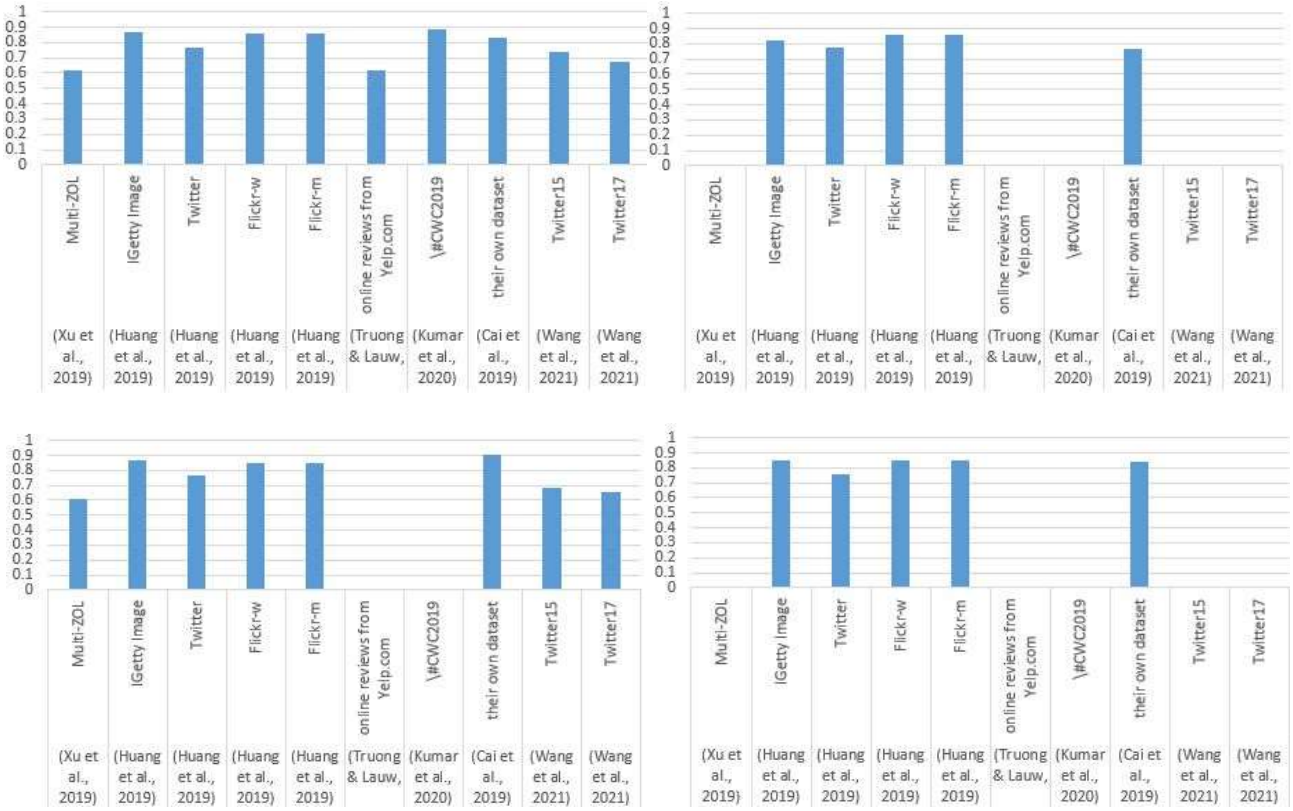
**Remark**

Table 2 compares these methods, and Fig. 1 compares their performances. With the support of massive amounts of data, DL-based methods for MSA far outperform machine learning-based MSA models. However, DL-based methods have their own limitations. For example, DL requires massive amounts of data (the small amount of data leads to the trained model overfitting). And DL still lacks interpretability (Zhong *et al.*, 2019). In practice, the long training time and high computational complexity of DL models are also major problems. Nevertheless, DL is still a popular method for processing MSA tasks.

Table 2: Cross-lingual pre-trained models for SA.

| Reference | Learning | Data set | Sentiment category |
|---|---|---|---|
| (Xu *et al.*, 2019) | Multi-Interaction Memory Network | Multi-ZOL | an integer sentiment score from 1 to 10 |
| (Huang *et al.*, 2019) | Deep Multimodal Attention Fusion | IGetty Image, Twitter, Flickr-w, Flickr-m | positive and negative. |
| (Truong & Lauw, 2019) | VistaNet | online reviews from Yelp.com | on the scale of 1 to 5 as five sentiment levels. |
| (Kumar *et al.*, 2020) | CNN based method | #CWC2019 | positive, negative, very positive, and very negative |
| (Cai *et al.*, 2019) | BiLSTM based method | their own dataset | Positive and negative |
| (Wang *et al.*, 2021) | EF-Net | Twitter15 and Twitter17 | Positive, negative, and neutral |

Source: This study.




Source: This study.

Figure 1: Performance comparison of various deep learning-based methods for static MSA.

## FUSION OF TEXT, IMAGE, AUDIO, AND VIDEO

This section discusses dynamic MSA on text, image, audio, and video. Some commonly used dataset for MSA are CMU-MOSI (http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/), MELD (https://affective-meld.github.io/), IEMOCAP (https://paperswithcode.com/dataset/iemocap), MOUD (http://multicomp.cs.cmu.edu/resources/moud-dataset/), CMUMOSEI (http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/), RAVDESS (https://smartlaboratory.org/ravdess/) and eNTERFACE'05 (http://www.enterface.net/results/).

**DeepCU**

To extract the common information from visual, text, and audio, Verma *et al.* (2019) propose an MSA model with universality and uniqueness, called DeepCU. Firstly, they use a deep convolutional tensor network to extract standard information from the multimodal representation. Secondly, they use a unique subnet to obtain modality-specific information. Finally, they integrate the two aspects of information through the fusion layer, which enhances the generalization performance of the MSA system. Their model outperformance existing technologies. In the future, they plan to introduce the attention network into the fusion layer to integrate the information obtained by the two subnets efficiently.

**LSTM based method**

To capture contextual information from their surroundings in the same video, Poria *et al.* (2017) analyze the context-related sentiments in user-generated videos. Specifically, they use textCNN to extract text features, openSMILE to extract audio features, 3D-CNN to extract visual features, and context-sensitive LSTM to extract the shared information among multimodal features. Their experiments show on datasets IEMOCAP, MOSI and MOUD, the accuracies are 76.1%, 80.3%, and 68.1%, respectively. Context-sensitive learning paradigm plays a key role in improving the model's performance. The model's performance on multimodal data (e.g., text-audio or text-video or text-audio-video) is better than that on single-modal data. For example, on dataset MOSI its accuracy is 78.12% on the monomodal data set of text, and its accuracy is 80.30% on the multimodal data set text-audio-video. Their experiments on the MOUD dataset show that visual information is more powerful than text and audio information.

**M3ER**

To learn to emphasize the more reliable cues and suppress others on a per-sample basis, Mittal *et al.* (2020) propose a learning-based model for MSA, called M3ER. It can use a multiplicative fusion layer to fuse cues from three co-occurring modalities of face, speech, and text. The layer can learn which modality should be relied on more for making a prediction. It is also more robust than other methods to sensor noise in any individual modality. This is because they introduce a modality check step that uses Canonical Correlational Analysis to distinguish ineffective modalities from effective modalities. Their experiments show that a mean accuracy of 82.7% on IEMOCAP and 89.0% on CMU-MOSEI, together meaning an improvement of about 5% over previous work.

**MA-RNN**

To fully use heterogeneous data, Kim and Lee (2020) propose a Multi-Attention Recurrent Neural Network (MA-RNN) for MSA on multimodal data of text, audio, and video. MA-RNN consists of two attention layers and a Bidirectional Gated Recurrent Neural Network (BiGRU). The first attention layer fuses data and reduces dimensionality. The second attention layer augments BiGRU to capture essential parts of the contextual information among utterances. In the feature extraction task, they use CNN to extract text features; use the open-source software openSMILE to extract audio features (such as pitch and voice intensity); use 3D-CNN to extract video features; and finally, use a fully connected layer to merge single-mode features. Their experiments show that MA-RNN achieves the state-of-the-art performance of 84.31% accuracy on the CMUMOSI dataset. Compared with the model without the attention mechanism, MA-RNN improves the accuracy by 17%.

**TransModality**

To reflect the information from the source modality and the target modality, the feature body can be learned and inspired by the recent success of Transformer in machine translation. Wang *et al.* (2020b) propose a fusion method, TransModality, for MSA. It is an end-to-end translation model with Transformer, which can mine the subtle correlation between modalities. Specifically, Transformer encodes the features of one modal and decodes it into features of another modal (the target modality) when outputting. They believe that sentimental polarity is related to not only mono-modal characteristics but also the connection between modalities. Using Transformer, the learned features contain the information from the source modality and the target modality. On multimodal datasets CMU-MOSI, MELD, and IEMOCAP, their experiments show that TransModality achieves state-of-art performance.

**Synch-Graph**

To emphasize the importance of constant interaction and co-learning between modalities, Mansouri-Benssassi and Ye (2020) propose an MSA method based on neural synchrony in multisensory audio-visual integration in the brain called Synch-Graph. Specifically, they use Spiking Neural Networks (SNN) to model the interaction between modalities, hoping to improve emotion classification accuracy. Many existing MSA methods focus on extracting features on each modality but ignore the importance of the interaction between modalities (Nian *et al.*, 2019; Zhang *et al.*, 2019). So, learning the interaction is the vital contribution of their model. They also use Graph Convolutional Networks (GCN) to learn neural synchrony patterns. Their experiments show that Synch-Graph achieves the overall accuracies of 98.3% and 96.82% on two state-of-the-art datasets,

respectively. In the future, it is worth examining the robustness of Synch-Graph on more massive data sets and studying the data representation problem after fusion.

**Unimodal Reinforced Transformer**

To explore the time-dependent interactions within unaligned sequences, He, Mai, and Hu (2021) proposed a unimodal reinforced Transformer with time squeeze fusion for MSA. In inferring emotions from language, sound, and visual sequences, previous research focused on analyzing aligned sequences, while unaligned sequence analysis is more practical in real scenes. Due to the long-term dependence hidden in the multimodal unaligned sequence and the lack of time alignment information, it is more challenging to explore the time-dependent interaction in the unaligned sequence. They introduced time squeeze fusion, which automatically explores time-related interactions by modeling unimodal and multimodal sequences from the perspective of compressed time dimensions. Their model achieved the most advanced performance in terms of accuracy and F1 score on the MOSEI dataset.

**Lightweight Deep Neural Networks**

To extract features of an audiovisual data stream that are fused in real-time for sentiment classification, Yakaew, Dailey, and Racharak (2021) propose a multimodal deep learning method. This method is suitable for automatic real-time analysis of emotions in the retail industry. They use modules based on small neural networks to dynamically content emotions from input video streams. They merge visual and auditory emotional characteristics and use fast-moving averages for emergencies, which become the final prediction. This article proves the use of lightweight predictive neural hierarchy. The network improves the feasibility of emotion classification performance based on multimodal processing in real-time on video data. The accuracy of this method in the Ryerson Emotional Voice Audiovisual Database (RAVDESS) reaches 90.74%.

**Channel-aware Temporal Convolution Network**

To reflect that the language modality is far more important than the acoustic and visual modalities, Mai, Xing, and Hu (2021) propose an MSA method based on a channel-aware temporal convolutional network with LSTM. Specifically, to enhance the language representation through the corresponding auxiliary modalities. Besides, they also introduced a "channel aware" time convolutional network to extract high-level representations of each mode to explore the interdependence between time and channel. Their experiments show that their method outperforms three widely used benchmarks of MSA.

**Multi-head Attention Mechanism**

To extract unimodal features and design a robust multimodal sentiment analysis model, Xi, Lu, and Yan (2020) proposed a method of extracting emotional features from vision, audio, and text and used these features to verify the MSA model based on a multi-head attention mechanism. Their experiments show that on datasets MOUDI and MOSI, its accuracy rates are 90.43% and 82.71%, respectively, meaning it is effective.

Table 3: Comparison of various deep learning-based methods for dynamic MSA.

| Reference | Learning | Data set | Sentiment category |
|---|---|---|---|
| (Verma *et al.*, 2019) | DeepCU | CMU-MOSI and POM | CMU-MOSI: positive and negative, POM: 16 sentiments |
| (Mahesh *et al.*, 2020) | BiLSTM | IEMOCAP28 and CMU-MOSI29 | IEMOCAP28: angry, happiness, neutral, and sadness; CMU-MOSI29: positive and negative. |
| (Zhang *et al.*, 2020c) | QT-LSTM | MELD, IEMOCAP | MELD: positive, negative, neutral, anger, disgust, fear, joy, neutral, sadness, surprise; IEMOCAP: anger, happiness, sadness, neutral. |
| (Harish & Sadat, 2020) | Attention-based DNN | CMU-MOSI | positive and negative |
| (Xi *et al.*, 2020) | Multi-head attention mechanism | MOUD, CMU-MOSI | Positive and negative |
| (Kim & Lee, 2020) | MA-RNN | CMU-MOSI | Positive and negative |
| (Mittal *et al.*, 2020) | M3ER:learning-based method | IEMOCAP, CMU-MOSEI | IEMOCAP: angry, happy, neutral, sad; CMU-MOSEI: anger, disgust, fear, happy, sad, surprise. |
| (Zadeh *et al.*, 2017) | Tensor fusion network | CMU-MOSI | 5-class and binary |
| (Zadeh *et al.*, 2018) | Dynamic fusion graph | CMUMOSEI | Anger, disgust, fear, happiness, sadness, and surprise. |
| (Mansouri - Benssassi & Ye, 2020) | Synch-Graph | eNTERFACE'05 and RAVDESS | Anger, disgust, fear, happiness, sadness, and surprise |

| | | | |
|---|---|---|---|
| (Yakaew *et* 2021) | Unimodal forced Transformer | RAVDESS | neutral, calm, happy, sad, angry, fearful, disgusted, and surprised |

**Remark**

Table 3 compares these methods, and Figure 2 compares their performances. Baltrusaitis *et al.* (2019) think that multimodal machine learning faces five significant challenges: (1) multimodal representation, (2) multimodal translation, (3) multimodal alignment, (4) multimodal fusion, and (5) multimodal co-learning. They summarise the state-of-art studies on each major challenge one by one. Multimodal machine learning is theoretically applicable to MSA tasks because the two face common challenges. In the future, it is worth thinking of whether or not there is a general multimodal machine learning algorithm that can do MSA. Besides, there are still many challenges with multimodal pre-training models in the era of deep learning. For example, compared with the NLP data set, the image and video data sets are smaller, but a pretraining model of this kind requires a massive amount of data. And the audio and video training data are different from the text. Audio and video have timing problems. According to a fixed time length, splitting the training data causes the training data to lack logical meaning. Also, the training costs are high.


Source: This study.

Figure 2: Performance comparison of various deep learning-based methods for dynamic MSA.

## MSA IN CONVERSATION

This section briefs MSA in conversations (a new and very challenging task). It aims to detect the sentimental state of speakers and track their sentimental changes during the dialogue. Some commonly used dataset for MSA in conversations are IEMOCAP (https://sail.usc.edu/iemocap/), SEMAINE (https://semaine-db.eu/), DailyDialog (http://yanran.li/dailydialog), sentimentLines (http://doraemon.iis.sinica.edu.tw/sentimentlines/index.html), EmoContex t(https://www.humanizing-ai.com/emocontext.html), MELD (https://affective-meld.github.io/) and MEISD (http://creativecommons.org/licenses/by/4.0/).

### CNN+LSTM-Based Method

To model the interactive information accurately, Zhang *et al.* (2020c) propose a comprehensive framework for MSA in multi-party conversations, called a Quantum-like Multimodal Network. Specifically, they firstly use a density matrix-based CNN to extract text and image features from conversations. Then, they input these features into an LSTM. Finally, they use the softmax function to obtain the result of SA. Their experiments show that their method significantly outperforms various baselines. Compared with LSTM, on the IEMOCAP and MELD data sets, the accuracy of their model is increased by 3.61% and 6.76%, respectively. The performance of their model largely depends on the representation quality of the density matrix. So, in the future, it is worth studying to capture more accurately interactive information between speakers.

### Memory Network-Based Method

To recognize utterance-level emotions in dyadic conversational videos, Hazarika *et al.* (2018) consider the contextual relation and dependency relation between discourses and use the dialogue memory network to solve the classification problem of video sentiment in conversation. They propose an architecture, termed CMN, for emotion detection in a dyadic conversation that considers utterance histories of both the speaker to model emotional dynamics. They perform experiments on the IEMOCAP dataset (Busso *et al.*, 2008). Their model is good at recognizing positive emotions, and when recognizing happiness and anger, its accuracy rates are 81.75% and 89.88%, respectively.

### RNN-Based Method

To leverage and capture the context of the conversation through all modalities, the dependency between the listener(s) and speaker emotional states, and the relevance and relationship between the available modalities, Shenoy and Sardana (2020) propose an end-to-end RNN based method for MSA in conversion. It considers the conversation context through all modalities, the dependency between the listener(s) and speaker emotional states, and the relevance and relationship between the available modalities. Their experiments show that their method outperforms the state-of-art method on a benchmark dataset in terms of accuracy, regression, and F1-score.

To capture the interlocutor state and contextual state between the utterances, Huddar, Sannakki, and Rajpurohit (2021a) propose a method based on RNN to obtain interlocutor state and context state between utterances. This article first extracts features from the text, sound, and visual features. An attention-based pairing technique is used to extract the context between utterances and to understand the relationship between forms and their importance before fusion. Their experiments show three standard data sets, IEMOCAP, CMU-MOSEI, and CMUMOSI, and their model is better than the standard baseline.

### Remark

SA in dialogues plays a critical role in dialogue data analysis (Wang *et al.*, 2020a). Although the single modal feature extraction methods employed by most researchers are similar, information fusion mechanisms are different, which can be summarised as feature-level fusion, decision-level fusion, and hybrid fusion mechanisms based on the two.

Unlike ordinary MSA tasks, MSA in conversation pays more attention to the interaction between different modalities. Here the interaction means the mutual influence of dialogue parties and joint representation and decision fusion between different modal data in a conversation. Inter-discourse interaction refers to the mutual influence caused by repeated interaction between dialogue parties.

For a human-machine dialogue system, since it is unavoidable that human has emotions in dialogue with the system, the system has to deal with these emotions. So, how to enable computers to understand and express sentiments in conversation is a new challenge in the fields of human-computer dialogue and SA.

## CONCLUSION

MSA is a critical issue in the natural language process. This paper surveys various MSA methods based on DL (including CNN, RNN, LSTM, Transformer, and Attention Machine). In particular, we analyze their advantages and disadvantages and point out the avenue for their improvement in the future. Besides, we review the work on MSA in conversation.

The audio and video data can reflect the user's tone. Hence, many researchers use speech data to help SA. One vital challenge in automatic speech SA is how to learn robust and discriminative representations for sentiment inferring. Besides, the use of computer vision to do SA is a relatively new topic. Its main task is to recognize and model sentiment information observed through faces, bodies, or gestures. The general direction of future work on MSA is data representation learning, feature extraction, and information fusion. The importance of the weight distribution of various modal data is also a topic worthy of

further discussion. In particular, what deserves MSA researchers focus on is the pre-trained models such as BERT and GPT-3. They are powerful, new paradigms of NLP and applicable to MSA. In (Macary *et al.*, 2021), the experimental pre-training model is used for speech emotion recognition.

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 61762016).

## REFERENCES

Abburi, H., Akkireddy, E. S. A., Gangashetti, S., & Mamidi, R. (2016). Multimodal sentiment analysis of telugu songs.In *Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016)*, IJCAI 2016, pages 48-52, New York City, USA.

Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey. *Information Fusion,* 76, 204-226. https://doi.org/10.1016/j.inffus.2021.06.003

Anastasopoulos, A., Kumar, S., & Liao, H. (2019). Neural language modeling with visual features. *arXiv preprint arXiv:1903.02930.*

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence,* 41(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation,* 42(4), 335-359. https://doi.org/10.1007/s10579-008-9076-6

Cai, Y., Cai, H., & Wan, X. (2019). Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2506-2515). https://doi.org/10.18653/v1/P19-1239

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics,* 9(3), 483. https://doi.org/10.3390/electronics9030483

Hamsa, S., Shahin, I., Iraqi, Y., & Werghi, N. (2020). Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier. *IEEE Access,* 8, 96994-97006. https://doi.org/10.1109/ACCESS.2020.2991811

Harish, A. B., & Sadat, F. (2020, April). Trimodal Attention Module for Multimodal Sentiment Analysis (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10), 13803-13804. https://doi.org/10.1609/aaai.v34i10.7173.

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L. P., & Zimmermann, R. (2018, June). Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, 2018, 2122. NIH Public Access. https://doi.org/10.18653/v1/n18-1193

He, J., Mai, S., & Hu, H. (2021). A Unimodal Reinforced Transformer With Time Squeeze Fusion for Multimodal Sentiment Analysis. *IEEE Signal Processing Letters,* 28, 992-996. https://doi.org/10.1109/LSP.2021.3078074

Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167, 26-37. https://doi.org/10.1016/j.knosys.2019.01.019

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification. *Computational Intelligence,* 36(2), 861-881. https://doi.org/10.1111/coin.12274

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. *International Journal of Multimedia Information Retrieval*, 9(2), 103-112. https://doi.org/10.1007/s13735-019-00185-8

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2021). Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN. *International Journal of Interactive Multimedia & Artificial Intelligence,* 6(6), 112-121. https://doi.org/10.9781/ijimai.2020.07.004

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2021). Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. *Multimedia Tools and Applications*, 80(9), 13059-13076. https://doi.org/10.1007/s11042-020-10285-x

Jazyah, Y. H. & Hussien, I. O. (2018). Multimodal Sentiment Analysis: A Comparison Study. *J. Comput. Sci.*, 14(6), 804-818. https://doi.org/10.3844/jcssp.2018.804.818

Ji, R., Cao, D., & Lin, D. (2015, April). Cross-modality sentiment analysis for social multimedia. In *2015 IEEE International Conference on Multimedia Big Data* (pp. 28-31). IEEE. https://doi.org/10.1109/BigMM.2015.85

Kaur, R. & Kautish, S. (2019). Multimodal sentiment analysis: A survey and comparison. International J*ournal of Service Science, Management, Engineering, and Technology (IJSSMET),* 10(2), 38-58. https://doi.org/10.4018/IJSSMET.2019040103

Kim, T. & Lee, B. (2020, June). Multi-attention multimodal sentiment analysis. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (pp. 436-441). http://dx.doi.org/10.1145/3372278.3390698

Kumar, A., Srinivasan, K., Cheng, W. H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management*, 57(1), 102141. https://doi.org/10.1016/j.ipm.2019.102141

Lin, P. & Luo, X. (2020). A survey of sentiment analysis based on machine learning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 12430, 372-387. Springer, Cham. https://doi.org/10.1007/978-3-030-60450-9_30

Lin, P. & Luo, X. (2020). A survey of the applications of sentiment analysis. *International Journal of Computer and Information Engineering*, 14(10), 334-346.

Lin, P., Luo, X., & Fan, Y. (2020). A Survey of Sentiment Analysis Based on Deep Learning. *International Journal of Computer and Information Engineering*, 14(12), 473-485.

Liu, B. & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_13

Macary, M., Tahon, M., Estève, Y., & Rousseau, A. (2021, January). On the use of Self-supervised Pre-trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)* (pp. 373-380). IEEE. https://doi.org/10.1109/SLT48900.2021.9383456

Mai, S., Xing, S., & Hu, H. (2021). Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 1424-1437. https://doi.org/10.1109/TASLP.2021.3068598

Mansouri-Benssassi, E., & Ye, J. (2020, April). Synch-graph: Multisensory emotion recognition through neural synchrony via graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(2), 1351-1358. https://doi.org/10.1609/aaai.v34i02.5491

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020, April). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(2), 1359-1367.. https://doi.org/10.1609/aaai.v34i02.5492

Nian, F., Chen, X., Yang, S., & Lv, G. (2019). Facial attribute recognition with feature decoupling and graph convolutional networks. *IEEE Access*, 7, 85500-85512. https://doi.org/10.1109/ACCESS.2019.2925503

Pérez-Rosas, V., Mihalcea, R., & Morency, L. P. (2013, August). Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1, 973-982.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. P. (2017, July). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics*, 1, 873-883. https://doi.org/10.18653/v1/P17-1081

Prabha, M. I. & Srikanth, G. U. (2019). Survey of sentiment analysis using deep learning techniques. In *2019 1st International Conference on Innovations in Information and Communication Technology* (ICIICT) (pp. 1-9). IEEE. https://doi.org/10.1109/ICIICT1.2019.8741438

Shenoy, A. & Sardana, A. (2020). Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation. *arXiv preprint arXiv:2002.08267*. doi:10.18653/v1/2020.challengehml-1.3

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65, 3-14. https://doi.org/10.1016/j.imavis.2017.08.003

Tenenbaum, J. B. & Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural computation*, 12(6), 1247-1283. https://doi.org/10.1162/089976600300015349

Truong, Q. T. & Lauw, H. W. (2019). Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 305-312. https://doi.org/10.1609/aaai.v33i01.3301305

Verma, S., Wang, C., Zhu, L., & Liu, W. (2019). Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization* (pp. 3627-3634). https://doi.org/10.24963/ijcai.2019/503

Wang, J., Gu, D., Yang, C., Xue, Y., Song, Z., Zhao, H., & Xiao, L. (2021). Targeted aspect based multimodal sentiment analysis: an attention capsule extraction and multi-head fusion network. *arXiv preprint arXiv:2103.07659.*

Wang, J., Wang, J., Sun, C., Li, S., Liu, X., Si, L., & Zhou, G. (2020). Sentiment classification in customer service dialogue with topic-aware multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5), 9177-9184. https://doi.org/10.1609/aaai.v34i05.6454

Wang, Z., Wan, Z., & Wan, X. (2020). Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020* (pp. 2514-2520). https://doi.org/10.1145/3366423.3380000

Xi, C., Lu, G., & Yan, J. (2020). Multimodal sentiment analysis based on multi-head attention mechanism. In *Proceedings of the 4th International Conference on Machine Learning and Soft Computing* (pp. 34-39). https://doi.org/10.1145/3380688.3380693

Xu, N., Mao, W., & Chen, G. (2019). Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 371-378. https://doi.org/10.1609/aaai.v33i01.3301371

Yakaew, A., Dailey, M. N., & Racharak, T. (2021). Multimodal Sentiment Analysis on Video Streams using Lightweight Deep Neural Networks. In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods* (pp. 442-451). https://doi.org/10.5220/0010304404420451

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., & Yang, K. (2020). Ch-sims: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3718-3727). https://doi.org/10.18653/v1/2020.acl-main.343

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018, July). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1, 2236-2246. https://doi.org/10.18653/v1/P18-1208

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017, September). Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1103-1114). https://doi.org/10.18653/v1/D17-1115

Zhang, C., Yang, Z., He, X., & Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3), 478-493. https://doi.org/10.1109/JSTSP.2020.2987728

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253. https://doi.org/10.1002/widm.1253

Zhang, M., Liang, Y., & Ma, H. (2019, July). Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 151-156). https://doi.org/10.1109/ICME.2019.00034

Zhang, Y., Song, D., Li, X., Zhang, P., Wang, P., Rong, L., ... & Wang, B. (2020). A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information Fusion*, 62, 14-31. https://doi.org/10.1016/j.inffus.2020.04.003

Zhong, Q., Fan, X., Luo, X., & Toni, F. (2019). An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications*, 117, 42-61. https://doi.org/10.1016/j.eswa.2018.09.038