Contents lists available at ScienceDirect

Information Fusion

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

# Multimodal sentiment analysis based on fusion methods: A survey

Linan Zhu [a], Zhechao Zhu [a], Chenwei Zhang [b], Yifei Xu [a], Xiangjie Kong [a,*]

[a] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
[b] School of Faculty of Education, University of Hong Kong, Hong Kong, China

ARTICLE INFO

*Keywords:*
Multimodal data
Sentiment analysis
Feature extraction
Fusion methods

ABSTRACT

Sentiment analysis is an emerging technology that aims to explore people's attitudes toward an entity. It can be applied in a variety of different fields and scenarios, such as product review analysis, public opinion analysis, psychological disease analysis, and risk assessment analysis. Traditional sentiment analysis only includes the text modality and extracts sentiment information by inferring the semantic relationship within sentences. However, some special expressions, such as irony and exaggeration, are difficult to detect via text alone. Multimodal sentiment analysis contains rich visual and acoustic information in addition to text, and uses fusion analysis to more accurately infer the implied sentiment polarity (positive, neutral, negative). The main challenge in multimodal sentiment analysis is the integration of cross-modal sentiment information, so we focus on introducing the framework and characteristics of different fusion methods. In addition, this article discusses the development status of multimodal sentiment analysis, popular datasets, feature extraction algorithms, application areas, and existing challenges. It is hoped that our work can help researchers understand the current state of research in the field of multimodal sentiment analysis, and be inspired by the useful insights provided in the article to develop effective models.

## 1. Introduction

### 1.1. From sentiment analysis to multimodal sentiment analysis

Sentiments are people's inherent attitudes toward a particular topic, person, or entity. Understanding people's attitudes is helpful for us to communicate, learn, and make decisions. For example, a company or store can make corresponding improvements based on how customers evaluate their brand or product. The evaluation of netizen voting can help government agencies guide public opinion. That is why, for the past two decades, AI researchers have been trying to give machines the cognitive ability to recognize, interpret, and express sentiments.

Early sentiment analysis mainly focused on text, in which only the interrelationships of words and phrases are considered to analyze sentiment [1]. However, relying on text data alone is not sufficient to extract the sentiments expressed by humans, because the meaning of what a speaker says often changes dynamically based on non-verbal behaviors. For example, the model's analysis of the word "great" in the text is generally positive; but if an exaggerated expression or sarcastic laughter is added, the expression may turn into a negative sentiment. Multimodal sentiment analysis has been proposed to address this problem, where *multimodal* refers to the multiple modalities (text, audio, and video) in which people communicate and express their feelings. In-depth research and work over the years have shown that multimodal systems are more effective than unimodal systems in identifying speakers' sentiments. A multimodal sentiment analysis survey published in 2015 reported that multimodal systems were consistently more accurate than their best unimodal counterparts [2].

With the tremendous development of social media, a large number of videos expressing personal opinions have been released on platforms such as YouTube or Facebook, which provides excellent resource support for multimodal sentiment analysis [3]. These videos are usually product reviews, movie reviews, policy critiques, etc. In addition to text information, videos also provide rich visual and acoustic information, and the feature fusion analysis of these modalities forms a multimodal sentiment analysis system [4].

The framework of a multimodal sentiment analysis system can be divided into the modeling of intra-modality dynamics and inter-modality dynamics. *Intra-modality dynamics* refers to the dynamical analysis of the interactions within each modality. For text, intra-modality dynamics are the interrelationships between words and phrases in a

[*] Corresponding author.
*E-mail address:* xjkong@ieee.org (X. Kong).

https://doi.org/10.1016/j.inffus.2023.02.028
Received 6 November 2022; Received in revised form 31 January 2023; Accepted 20 February 2023
Available online 24 February 2023

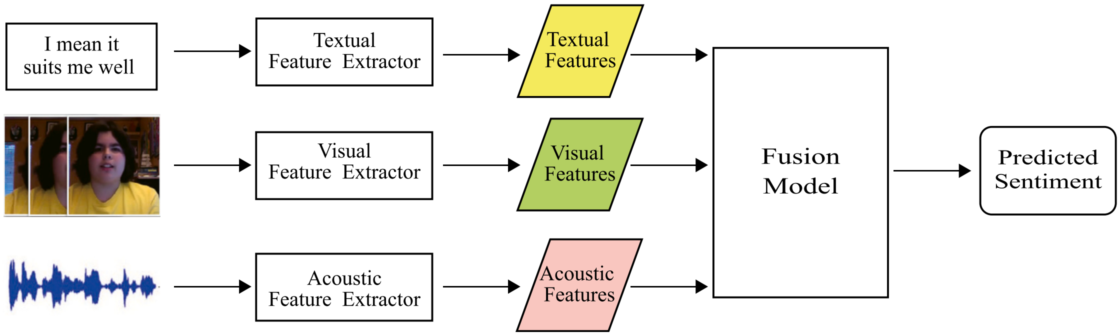

**Fig. 1.** General framework of a typical multimodal sentiment analysis system.

sentence. On the other hand, *inter-modality dynamics* refers to the interaction between different modalities, which can be further divided into synchronous and asynchronous according to whether the modalities are aligned or not. Synchronous inter-modal interaction means that the modalities have been aligned, and different modalities with the same timestamp appear simultaneously. For example, when a speaker expresses his opinion on the screen, the text can be well-matched. Asynchronous inter-modal interaction means that there is no alignment between modalities, and different modalities may not appear simultaneously. For example, the speaker's voice may be heard after they are shown speaking. Existing datasets are often unaligned but can be aligned using alignment tools such as P2FA [5]. Since the intra-modality representation must be obtained accurately and the interaction between modalities is very complex, the biggest challenge of multimodal sentiment analysis is how to create the intra-modality representation and find the best fusion method to explore the inter-modality interaction.

### 1.2. Scope of this survey

As more and more articles on multimodal sentiment analysis are published, survey papers are needed to summarize the latest research methods and forecast future research trends in this field. As early as 2017, Poria et al. [6] proposed a review of affective computing from unimodal to multimodal fusion. Their survey elaborates on some basic feature extraction methods and model frameworks for emotion recognition and sentiment analysis. In the same year, Soleymani et al. [7] also summarized the current situation of multimodal sentiment analysis and offered an outlook on existing applications and future development trends. Although these survey papers provided a comprehensive overview of the current development at that time, many rich datasets and advanced models have been proposed in the past few years. For example, two of the most popular datasets in the field (CMU-MOSI [8] and CMU-MOSEI [9]) are not present in these earlier surveys; nor are models based on attention mechanisms.

In 2021, Gkoumas et al. [10] conducted detailed experiments and analyses on 11 state-of-the-art models using CMU-MOSI and CMU-MOSEI. The authors found that models that use attention mechanisms often achieve better results. However, the number of experimental models in this paper is too small to offer a more detailed generalization about the field. Chandrasekaran et al. [11] investigated the application of multimodal sentiment analysis to social media and proposed a large number of methods and applications. Based on the modules used by the models, Abdu et al. [12] divided 35 models into 8 categories and gave an overview. However, none of these survey papers provide a detailed description of the models that exist in the field from the perspective of fusion methods.

Our work focuses on the sentiment analysis of three modalities (text, audio, and video) and does not address other tasks such as bimodal sentiment analysis or emotion recognition. According to the process of multimodal sentiment analysis, we first list some popular datasets in the field and analyze the commonly used feature extraction methods for each modality. Then the fusion forms of existing models are analyzed, and a taxonomy framework of fusion methods with eight classifications is established. The fields where multimodal sentiment analysis can be applied and the main challenges of the current model will be given later. Finally, we summarize the content and contributions of the full paper and illustrate several possible future research trends.

Our work aims to:

1. Provide an overview of existing work which will help researchers gain a detailed understanding of available methods and resources for multimodal sentiment analysis.

2. Classify existing model frameworks from the perspective of fusion methods and give detailed descriptions of each method.

3. Summarize the application fields, expound on the existing challenges and identify future research directions.

### 1.3. Multimodal sentiment analysis process

Fig. 1 shows the general framework of a typical multimodal sentiment analysis system. The framework can be divided into two parts: unimodal data processing and multimodal data fusion. First, feature extractors are applied to textual, visual, and acoustic data respectively to extract features. Then, the extracted features are transferred to the fusion model to predict sentiment. Both of these components are important for the performance of the whole model. The poor unimodal analysis leads to an insufficient understanding of intra-modal interactions and degrades the performance of multimodal systems; inefficient multimodal fusion leaves the interaction between modalities incompletely utilized, which affects the stability of the multimodal system.

The paper is organized as follows: Section 2 summarizes the most popular datasets in multimodal sentiment analysis. Section 3 discusses commonly used feature extraction techniques and related articles. Section 4 classifies advanced models in multimodal sentiment analysis into eight categories according to their fusion methods, and analyzes their advantages and disadvantages while introducing the models in detail. Section 5 discusses possible applications, and Section 6 illustrates some of the challenges of existing models. Finally, the conclusions and future research trends are discussed in Section 7.

## 2. Popular datasets in multimodal sentiment analysis

Table 1 presents popular datasets in the field of multimodal sentiment analysis. The first three columns show, respectively, the name, publication year, and number of videos in the dataset. The fourth column is the task type targeted by the dataset, including review videos, news videos, and movies. The fifth column shows the language used in the dataset, and the sixth column shows the sources of the dataset, mostly mainstream social media and some movies. The seventh column is the sentiment label annotated by the dataset, represented by the

**Table 1**
Statistics on popular multimodal sentiment analysis datasets.

| Dataset | Year | Videos | Task | Language | Source | Sentiments | Available at |
|---|---|---|---|---|---|---|---|
| YouTube [4] | 2011 | 47 | Review | English | YouTube | [−1, +1] | http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/ |
| MOUD [13] | 2013 | 80 | Review | Spanish | YouTube | [−1, +1] | http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/ |
| ICT-MMMO [14] | 2013 | 370 | Review | English | YouTube, ExpoTV | [−2, +2] | http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/ |
| NRS [15] | 2014 | 929 | News | English | American news programs and channels | [−1, +1] | https://www.ee.columbia.edu/ln/dvmm/newsrover/sentimentdataset/ |
| POM [16] | 2016 | 1000 | Review | English | ExpoTV | [1, 7] | http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/ |
| CMU-MOSI [8] | 2016 | 93 | Review | English | YouTube | [−3, +3] | http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/ |
| CMU-MOSEI [9] | 2018 | 3228 | Review | English | YouTube | [−3, +3] | http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/ |
| CH-SIMS [17] | 2020 | 60 | Movie | Chinese | movies, TV series, variety shows | [−2, +2] | https://github.com/thuiar/MMSA |
| CMU-MOSEAS [18] | 2020 | 4000 | Review | Spanish, German, Portuguese, French | YouTube | [−3, +3] | https://bit.ly/2Svbg9f |

Likert scale. Specifically, [−1,+1] corresponds to [−1: negative, 0: neutral, +1: positive]. [−2,+2] corresponds to [−2: strongly negative, -1: weakly negative, 0: neutral, +1: weakly positive, +2: strongly positive]. [−3,+3] likewise represents a 7-point scale from highly negative to highly positive. [1,7] corresponds from 1 (very unpersuasive) to 7 (very persuasive). The last column provides access to the datasets, some of which can be downloaded directly from the website, and some of which can be obtained by email to the authors.

### 2.1. YouTube dataset

The YouTube dataset was developed by Morency et al. [4] in 2011 and consists of 47 videos collected from YouTube. The dataset contains three modalities: visual, textual, and acoustic, and is the first dataset to be used in the trimodal sentiment analysis task. These videos are characterized by diversity, multimodality, and the presence of ambient noise. The authors collected the videos not based on a specific topic, but via the following keywords: opinion, product review, best perfume, toothpaste, war, job, business, camera review, I hate... etc. 47 videos are divided into 20 female and 27 male speakers who are approximately 14–60 years old. Although the speakers come from different cultural backgrounds, they all express themselves in English. Each video contains 3–11 utterances and is assigned one of three labels: negative, neutral, or positive.

### 2.2. MOUD dataset

Developed by Perez-Rosas et al. [13] in 2013, the Multimodal Opinion Utterances Dataset (MOUD) is the first multimodal opinion database annotated at the discourse level, addressing the relative importance of modality and individual characteristics. The authors collected a set of videos of product opinions expressed in Spanish from YouTube, using the following keywords: my favorite products, not recommended products, not recommended movies, recommended books, etc. The keywords are not specific to a particular product type; instead, a wide variety of product names are included, so the dataset has a degree of generality across the broad field of product reviews. The dataset consists of 80 videos randomly selected from the collected videos, including 15 male speakers and 65 female speakers aged approximately 20 to 60. A 30-second opinion segment is manually selected from each video and then split into an average of 6 utterances, resulting in a dataset of 498 utterances. The average duration of these utterances is 5 s, with a standard deviation of 1.2 s. Each utterance is labeled as positive, negative, or neutral.

### 2.3. ICT-MMMO dataset

Developed by Wollmer et al. [14] in 2013, the Institute for Creative Technologies' Multi-Modal Movie Opinion (ICT-MMMO) dataset includes a collection of real review videos from YouTube and ExpoTV, which mainly contains movie review videos by non-professional users. Of the 370 movie review videos collected, 228 were positive reviews, 23 were neutral, and 119 were labeled negative. Each reviewer speaks in English, and commentary videos vary in length from 1–3 min. The authors followed previous work on sentiment analysis and used 5 sentiment labels: strongly negative, weakly negative, neutral, weakly positive, and strongly positive.

### 2.4. NRS dataset

Developed by Ellis et al. [15] in 2014, the News Rover Sentiment (NRS) dataset is the first dataset to study sentiment analysis in the field of news. The authors collected videos of various US news programs and channels recorded between August 13, 2013, and December 25, 2013. The video length of the dataset is limited to between 4 and 15 s. The reason for this restriction is that the authors believe that short videos make it difficult to truly decipher people's emotional expressions, and videos longer than 15 s may contain multiple sentences with different sentiments. The final dataset has 929 clips, each annotated with three categories of sentiment: positive, negative, or neutral.

### 2.5. POM dataset

The Persuasive Opinion Multimedia (POM) dataset, developed by S. Park et al. [16] in 2016, includes persuasive subjective annotations and high-level related attributes. The dataset includes 500 5-star review videos (306 men, 194 women) and 500 1- or 2-star review videos (363 men, 137 women) collected from ExpoTV. A 5-star rating is considered positive, and a 1- or 2-star rating is considered negative. The average length of the videos is about 93 s, with a standard deviation of about 31 s. The authors annotated the persuasiveness of the speaker from 1 (very unconvincing) to 7 (very convincing). In addition to investigating persuasiveness, another feature of the dataset is a better understanding of other high-level attributes that may be associated with persuasiveness (e.g., confidence, trustworthiness, dominance, humor, passion, physical attractiveness, and professional appearance). The authors argued that additional annotation of high-level attributes will make the dataset more widely applicable to other related research topics (such as personality trait modeling).

### 2.6. CMU-MOSI dataset

Developed by Zadeh et al. [8] in 2016, the Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) dataset is the first opinion-level annotation corpus for sentiment and subjectivity analysis in online video. Besides annotating subjectivity and sentiment intensity, visual features of each opinion annotation and audio features annotated every millisecond were also included. The MOSI dataset contains a total of 3702 video clips, including 2199 opinion clips. The sentiment of each opinion section is annotated as a range from highly negative to highly positive. The final dataset includes a total of 93 randomly selected videos featuring 89 different speakers. It is worth mentioning that visual features such as facial action units and over 32 audio features have been automatically extracted from MPEG files.

### 2.7. CMU-MOSEI dataset

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset, developed by Zadeh et al. [9] in 2018, is the next generation of CMU-MOSI. CMU-MOSEI contains 23,453 annotated video clips from 1000 different speakers (57% male and 43% female). During data collection, the authors used face detection to analyze whether there is only one speaker in the video to ensure the video is a monologue. The dataset has a total of 250 topics: the 3 most common topics are reviews (16.2%), debate (2.9%), and consulting (1.8%) while the remaining topics are almost evenly distributed. Sentiment annotation is the same as in CMU-MOSI: each sentence is annotated for a sentiment on a [−3,3] Likert scale. Ekman's emotions of happiness, sadness, anger, fear, disgust, and surprise are also annotated on a [0,3] Likert scale for the presence of emotion x: [0: no evidence of x, 1: weakly x, 2: x, 3: highly x].

### 2.8. CH-SIMS dataset

The Chinese Single- and Multimodal Sentiment (CH-SIMS) dataset was developed by Yu et al. [17] in 2020. The authors collected 60 original videos from movies, TV series, and variety shows, and obtained 2281 video clips (1500 male, 781 female). The characters in these video clips have rich personality backgrounds and a wide age range, and each video clip has multimodal annotations and three modality-independent unimodal annotations. Researchers can both study interactions between modalities and perform unimodal sentiment analysis using independent unimodal annotations. The average length of segments in SIMS is 3.67s, and the average word length of each clip is 15. The annotation process started with unimodal annotation for each tag, in the order of text, audio, silent video, and finally multimodal annotation. For each clip, the authors classified its sentiment into one of five categories: negative, weakly negative, neutral, weakly positive, and positive.

### 2.9. CMU-MOSEAS dataset

After the work of CMU-MOSI and CMU-MOSEI, considering that some widely spoken languages have few or no large-scale datasets in the field of multimodal sentiment analysis, Zadeh et al. [18] proposed the CMU Multimodal Opinion Sentiment, Emotions and Attributes (CMU-MOSEAS) dataset in 2020. This is the first large-scale multimodal language dataset for Spanish, Portuguese, German, and French. Videos were found manually from YouTube using more than 250 different search terms, and no more than five videos were collected from a single channel to ensure diversity across speakers. The authors annotated the sentences with 20 labels, including sentiment, subjectivity, emotions, and attributes, where sentiment was annotated as [−3,+3]. The entire dataset contains a total of 4000 monologue videos (1000 per language), spanning 1645 speakers and 40 000 annotated sentences (10 000 per language). Each language has a large set of unsupervised sentences to enable unsupervised pre-training of multimodal representations. CMU-MOSEAS places special emphasis on protecting the privacy of speakers. Although the videos are publicly available on YouTube, a specific EULA (End User License Agreement) is still required to download the labels. A unique aspect of CMU-MOSEAS is that it allows multimodal statistical comparisons between different languages; for example, Spanish and Portuguese report positive sentiment more commonly.

## 3. Feature extraction

Unimodal feature extraction is an important building block for multimodal sentiment analysis systems. In this section, we will introduce the feature extraction methods of text, video, and audio respectively, and list some works using the feature extraction methods we mentioned, as shown in Table 2. To better provide inspiration, we only focus on the popular methods since deep learning was widely developed.

**Table 2**
The feature extraction methods used in models.

| Model | Textual | Visual | Acoustic |
|---|---|---|---|
| THMM [4] | Polarized words | OKAO Vision | OpenEAR |
| SVM [13] | Bag-of-words | CERT | OpenEAR |
| MKL [19] | Word2vec | CLM-Z | openSMILE |
| SAL-CNN [20] | Word2vec | CLM-Z | openSMILE |
| TFN [21] | GloVe | Facet | COVAREP |
| LMF [22] | GloVe | Facet | COVAREP |
| HFFN [23] | GloVe | Facet | COVAREP |
| LMFN [24] | GloVe | Facet | COVAREP |
| GME-LSTM(A) [25] | GloVe | Facet | COVAREP |
| MARN [26] | GloVe | Facet | COVAREP |
| MFN [27] | GloVe | Facet | COVAREP |
| RAVEN [28] | GloVe | Facet | COVAREP |
| SWRM [29] | BERT | Facet | COVAREP |
| MCTN [30] | GloVe | Facet | COVAREP |
| MulT [31] | GloVe | Facet | COVAREP |
| MAG [32] | BERT | Facet | COVAREP |
| ICDN [33] | GloVe | Facet | COVAREP |
| AMOA [34] | BERT | OpenFace 2.0 | openSMILE |
| ICCN [35] | BERT | Facet | COVAREP |
| MISA [36] | BERT | Facet | COVAREP |
| HyCon [37] | BERT | Facet | COVAREP |
| HGraph-CL [38] | BERT | Facet | COVAREP |
| BC-LSTM [39] | Text-CNN | 3D-CNN | openSMILE |
| MMMU-BA [40] | GloVe | Facet | COVAREP |

### 3.1. Textual feature extraction

With the development and maturation of deep learning technologies such as neural networks, word embedding technology has been applied to the field of text feature extraction. Word embedding uses neural networks to learn the correlation between parts of speech and word meanings, and expresses words with similar meanings in the form of vectors with close Euclidean distances. It can convert high-dimensional sparse vectors into low-dimensional dense vectors, which mitigates computational demands and solves the problem that vectors do not contain spatial and semantic information. Common word embedding methods include NNLM [41], HLBL [42], and Word2Vec [43], with the latter the most commonly used. Word2Vec contains two different styles of models: CBOW and Skip-gram. These have different goals in that the former predicts the center word based on the surrounding words, while the latter does the opposite. These methods can capture complex patterns beyond lexical similarity, with general improvements on other tasks as well. However, the disadvantages are that large datasets are required and statistics cannot be fully utilized.

Most recent work uses GloVe to extract text features. In addition, large pre-trained models such as BERT are often used. GloVe [44] shares many conceptual similarities with Word2Vec and adds statistics-based information. This enables GloVe to use both the global information of the corpus and local contextual features. BERT can process the entire sequence in parallel, using an attention mechanism to gather contextual information about words. It is then encoded with a rich vector representing the context so that words related to all other words in the sentence are processed simultaneously. The model can learn how to derive the meaning of a given word from other words in the sentence.

### 3.2. Visual feature extraction

Visual feature extraction serves mainly to extract people's facial expression features and body posture from video because the information contained in it is the key to analyzing the speaker's sentiment. Some networks (especially CNN) have a good ability to extract features from images, avoiding the tediousness of manual feature extraction. Benitez-Quiroz et al. [45] proposed a deep neural architecture that addresses this problem by combining learned local and global features in the initial stage and replicates the message passing algorithm between classes, similar to the graphical model inference approach in later stages. Tran et al. [46] proposed a simple and efficient deep three-dimensional convolutional neural network (3D-CNN) for spatiotemporal feature extraction, which can be used in different tasks, such as action recognition, same action judgment, and dynamic scene recognition.

Nowadays, most of the features are extracted by neural networks or public libraries. The most commonly used public libraries include OKAO, CERT, OpenFace, and Facet. OKAO Vision detects and extracts facial features at each frame, then returns a smile intensity (0–100) and eye gaze direction. Computer expression recognition toolbox (CERT) [47] allows users to automatically extract the following visual features: smile and head pose estimates, facial AUs, and eight basic emotions (anger, contempt, disgust, fear, joy, sadness, surprise, and neutral). The MultiComp OpenFace 2.0 toolkit [48] extracts 68 facial landmarks, 17 facial action units, head pose, head orientation, and eye gaze. The Facet library extracts a set of visual features, including facial action units, facial landmarks, head pose, gaze tracking, and HOG features.

### 3.3. Acoustic feature extraction

Deep learning has also attracted more and more attention in audio classification research. Long Short-Term Memory (LSTM) [49] and bidirectional LSTM [50] have been widely used for hand-extracted acoustic features. Since deep networks are often used to automatically extract features in computer vision, a research question is whether the network can be replicated or not. The answer was given by Anand et al. [51], who used CNN to extract features from audio and then passed them into the classifier for the sentiment classification task. Deep neural networks based on generalized discriminant analysis (GDA) are also very popular for automatic feature extraction from raw audio data.

Recently, most multimodal sentiment analysis models use OpenEAR, openSMILE, LibROSA, COVAREP, and other open-source libraries to extract acoustic features. The open source software OpenEAR [52] automatically computes a set of acoustic features, including prosody, energy, vocal probability, spectrum, and cepstral features, and uses z-standardization for speaker normalization. All features and functions can be calculated using the online audio analysis toolkit openSMILE [53]. Specifically, the features extracted by openSMILE consist of several low-level descriptors (LLDs) such as MFCC, pitch, and sound intensity and their statistical functions. Some of these functions are amplitude mean, arithmetic mean, root quadratic mean, standard deviation, etc. The LibROSA speech toolkit [54] can be used to extract acoustic features at 22 050 Hz. A total of 33-dimensional frame-level acoustic features are extracted, including 20-dimensional MFCC and 12-dimensional constant q transform (CQT). These traits are related to mood and speech intonation. For each opinion utterance audio, the COVAREP acoustic analysis framework [55] can also be used to extract acoustic features, including 12 MFCC, glottal source parameters, peak slope parameters, maximum dispersion quotients (MDQ), and Liljencrants-Fant (LF).

## 4. Fusion methods

The use of efficient methods to fuse feature information from different modalities is a major challenge for multimodal sentiment analysis. In this section, we classify 42 methods into 8 categories according to their fusion methods, as shown in Fig. 2. We describe the framework of each model in detail and list the advantages and disadvantages of each model, which can provide inspiration for readers to carry out their work. At the end of this section, we systematically compare the fusion methods of each classification and describe the development driven by the advantages and disadvantages of the models.

### 4.1. Early fusion

Early fusion is also called feature-level fusion. By extracting the features of each modality and merging them at the input level, a joint representation is constructed, and sentiment classification is performed on this basis. The framework of this approach can be simple as it relies on general models (Support Vector Machine (SVM [56]) or deep neural networks) to learn view-specific and cross-view dynamics without any specific model design. However, early-stage fusion results in a lack of detailed modeling of view-specific dynamics, thereby losing contextual and temporal dependencies within each modality, which in turn affects the modeling of cross-view dynamics and leads to an overfitting of the data. Table 3 summarizes the advantages and disadvantages of each model.

#### 4.1.1. THMM (Tri-modal Hidden Markov Model)

Morency et al. [4] first proposed the task of tri-modal sentiment analysis and designed a model to solve it. After extracting the features of each modality and concatenating them, a tri-modal HMM classifier is used to learn the hidden structure of the input signal.

#### 4.1.2. SVM (Support Vector Machine)

Perez-Rosas et al. [13] combined the features collected from all multimodal streams into a feature vector, thereby generating a vector for each utterance and using an SVM classifier to decide the sentimental orientation of the utterance. S. Park et al. [16] used Support Vector Machines (SVMs) for classification and Support Vector Regressions (SVRs) for regression experiments with the radial basis function kernel as the prediction models.

#### 4.1.3. MKL (Multiple Kernel Learning)

Poria et al. [19] used two different feature selectors to reduce the number of features after extracting tri-modal features. One is based on the cyclic correlation-based feature subset selection (CFS), and the other is based on principal component analysis (PCA). In addition to improving the processing time of the model, the feature selection method also slightly improves the experimental results. Finally, the processed feature vectors are concatenated, and the classifier is trained with a multi-kernel learning (MKL) algorithm. In the following year, based on previous work, the authors [57] proposed a convolutional recurrent multiple kernel learning (CRMKL) model, specifically using a convolutional RNN for visual sentiment detection, which further improved the experimental results.

#### 4.1.4. EF-LSTM (Early Fusion LSTM)

Zadeh et al. [26] concatenated the inputs from different modalities at each time step and used it as the input to a single LSTM.

#### 4.1.5. Self-MM (Self-Supervised Multi-task Multimodal sentiment analysis network)

Yu et al. [58] proposed a Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) to divide the multimodal sentiment analysis task into a multimodal task and three independent unimodal subtasks. The multimodal task is similar to other early fusion methods, which is why this method is classified in the category of early fusion methods. A major feature of Self-MM is the design of a label generation module based on a self-supervised learning strategy to obtain independent unimodal supervision. For example, if the multimodal annotation is closer to the positive center and the unimodal representation is more negatively centered, the Unimodal Label Generation Module (ULGM) will impose a negative relative offset value to the multimodal label to form the Unimodal Label. A hard sharing strategy is adopted to share the bottom representation learning network between the multimodal task and different unimodal tasks. The result is that the multimodal and unimodal tasks are jointly trained to learn their congruence and dissimilarity, respectively. The experiment results verify the reliability and stability of the automatically generated unimodal labels.

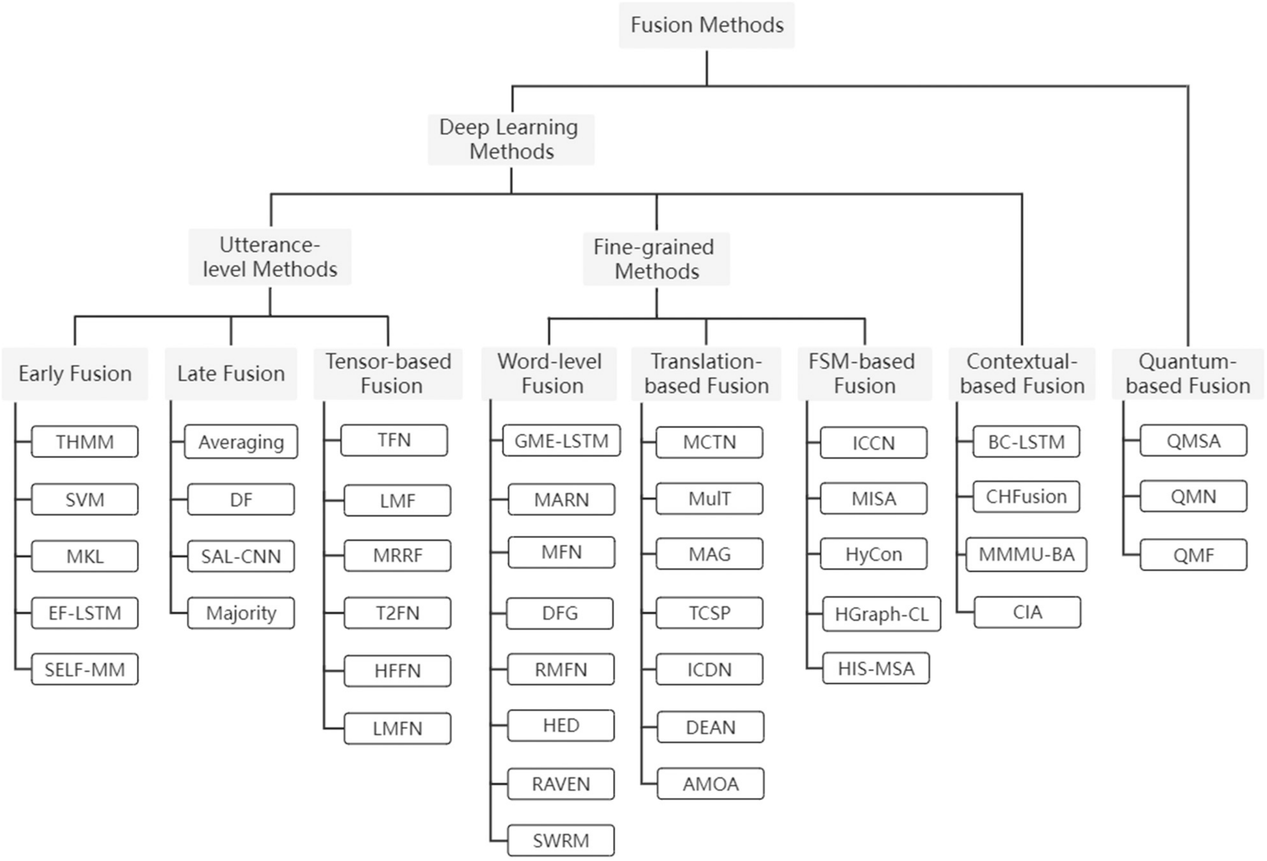

**Fig. 2.** Effective multimodal fusion models in multimodal sentiment analysis.

**Table 3**
Characteristics of early fusion methods.

| Methods | Advantages | Disadvantages |
|---|---|---|
| THMM | • The task of trimodal sentiment analysis is proposed for the first time<br>• Tri-modal Hidden Markov Model (THMM) | • Small scale dataset |
| SVM | • Support Vector Machines (SVM) | • Difficult to implement for large-scale training samples |
| MKL | • Multiple kernel learning (MKL)<br>• Use deep CNN to extract features from text | • Small datasets can lead to early overfitting |
| EF-LSTM | • Concatenate the inputs from different modalities at each time step and use that as the input to a single LSTM | • The training process is difficult to parallelize |
| SELF-MM | • A unimodal label generation module based on the self-supervised strategy<br>• A novel weight self-adjusting strategy is introduced to balance different task loss constraints | • The generated audio and vision labels are limited by the preprocessing features |

### 4.2. *Late fusion*

Late fusion, also known as decision-level fusion, first conducts sentiment analysis based on each modality, and then proposes different mechanisms to incorporate unimodal sentiment decisions into the final decision, including averaging [59], majority voting [60], weighted sum [61], or learnable models. This fusion approach is usually strong in modeling view-specific dynamics, and due to the integration of its modules, the model is usually lightweight, flexible, and adapts well to changes in the number of modalities. However, since separate models are built for each modality, inter-modal interactions are often not modeled efficiently because the dynamics between these modalities are often more complex than decision voting. Table 4 summarizes the advantages and disadvantages of each model.

#### 4.2.1. *Averaging*

Nojavanasghari et al. [62] trained a unimodal classifier for each of the three modalities, then averaged the confidence scores of individual unimodal classifiers to make the final prediction.

#### 4.2.2. *DF (Deep Fusion)*

Nojavanasghari et al. [62] trained a unimodal classifier for each of the three modalities, then used the confidence score of each unimodal classifier (c); along with the complementary scores (1-c) as input features to a fusing deep network that was used to make final predictions.

#### 4.2.3. *SAL-CNN (Select-Additive Learning CNN)*

Wang et al. [20] proposed a SAL-CNN model. After the CNN model is fully trained, the authors use SAL to improve its generality and predict sentiment, trying to prevent identity-dependent information from

**Table 4**
Characteristics of late fusion methods.

| Methods | Advantages | Disadvantages |
|---|---|---|
| Averaging | • Average the output of unimodal classifiers | • If a modality is noisy, it may affect the final prediction |
| DF | • The model learns the importance of each modality and assigns weights to the final model accordingly | • Cannot handle noisy training data |
| SAL-CNN | • Select-Additive Learning<br>• Force the original model to discard confounding elements to improve the generality of the trained neural network | • Added Gaussian noise may impair normal model predictions |
| Majority | • Do majority voting for classification tasks and predict expected labels for regression tasks | • Some close key outputs may be ignored |

**Table 5**
Characteristics of tensor-based fusion methods.

| Methods | Advantages | Disadvantages |
|---|---|---|
| TFN | • Learn both the intra-modality and inter-modality dynamics end-to-end<br>• The tensor fusion layer uses a 3-fold Cartesian product of the three output vectors of the embedding layer | • The resulting representation has a very large dimensionality and thus a large number of parameters |
| LMF | • Similar to TFN, but adds an additional low-rank factor to reduce computational memory<br>• Can perform robustly in a wide range of low-rank settings and be more efficient in training and inference | • Modeling of local interactions neglected |
| MRRF | • Tuckers tensor decomposition method allows different compression ratios for each modality<br>• Redundant information repeated across modalities can be eliminated and lead to less information loss and minimal parameters | • Require training and testing on larger datasets |
| T2FN | • Temporal Tensor Fusion Network (T2FN)<br>• Regularization method based on tensor rank minimization | • Imperfect data increases tensor rank |
| HFFN | • Divide, Conquer and Combine<br>• Local fusion by exploring the interaction of partially aligned feature vectors of different modalities within a sliding window<br>• Explore the attention mechanism between local interactions through an Attentive Bi-directional Skip-connected LSTM | • It may not be optimal for all three modalities to be treated equally<br>• There may be a lot of redundant information in the feature vector |
| LMFN | • The feature vector corresponding to each modality is divided into multiple segments, and each local interaction is learned through a tensor fusion procedure<br>• Model global interactions using a Bidirectional Multiconnected LSTM<br>• The calculation cost is linearly related to the dimension of the feature vector | • Performance on CMU-MOSEI is weaker than on other datasets<br>• Dividing the feature vectors into segments of equal size may not be optimal |

being learned in a deep neural network. The SAL method consists of a two-stage process (selection and addition). During the selection phase, SAL identifies the confounding factors from the latent representation learned by neural networks. During the addition phase, SAL forces the original model to discard the confounding elements by adding Gaussian noises to these representations. SAL nearly maintains the clustering structure of identity, but greatly improves the clustering structure of the category of sentiment.

#### 4.2.4. Majority

Zadeh et al. [26] performed majority voting for classification tasks, and predicted the expected label for regression tasks.

### 4.3. Tensor-based fusion

Tensor-based methods mainly compute the tensor product of unimodal sentence representations to obtain multimodal sentence representations. This requires first converting the input representation into a high-dimensional tensor, then mapping it back to a low-dimensional output vector space, which is a typical non-concatenated feature fusion method. Tensors are powerful because they capture important higher-order interactions across time, feature dimensions, and multiple modalities [63]. However, a drawback of this approach is that the computational complexity grows exponentially and there is no fine-grained word-level interaction between cross-modalities. The method first embeds the three modalities of text, vision, and audio to explore the dynamics within the modalities, then fuses multimodal embedding representations to explore dynamic interactions between modalities. Table 5 summarizes the advantages and disadvantages of each model.

#### 4.3.1. TFN (Tensor Fusion Network)

Zadeh et al. [21] proposed a Tensor Fusion Network (TFN) model, which learns both the intra-modality and inter-modality dynamics end-to-end. The model is divided into three parts: Modality Embedding Subnetworks, Tensor Fusion Layer, and Sentiment Inference Subnetwork.

The Modality Embedding Subnetworks use an LSTM network with forget gates to learn time-dependent language representations, which are then joined to a fully connected network to obtain language embeddings. For acoustic and visual features, FACET and COVAREP are used to extract features, and after average pooling, they are respectively connected to deep neural networks to obtain embeddings. In the Tensor Fusion Layer, a triple Cartesian product is used for the three output vectors of the embedding layer, which fully combines the unimodal, bimodal, and trimodal interactions in tensor fusion. The resulting multimodal tensor is passed to a fully connected deep neural network called the Sentiment Inference Subnetwork, and the prediction result is obtained.

#### 4.3.2. LMF (Low-rank Multimodal Fusion)

To address the problem of exponential increase in computational complexity introduced when converting inputs to tensors, Liu et al. [22] proposed a Low-rank Multimodal Fusion (LMF) method, which uses low-rank tensors to perform multimodal fusion, greatly reducing the computational complexity. The model is similar to TFN, but decomposes the weights into low-rank factors, reducing the number of parameters. Moreover, the experimental results show that the proposed model can perform robustly in a wide range of low-rank settings, and it is more effective in training and reasoning than other methods using tensor representation.

#### 4.3.3. MRRF (Modality-based Redundancy Reduction Fusion)

Inspired by TFN [21] and LMF [22], Barezi et al. [64] proposed a modality-based redundancy reduction fusion (MRRF) model to understand and modulate the relative contribution of each modality in multimodal inference tasks. Whereas the factorization in LMF utilizes a single compression rate across all modalities, MRRF uses Tucker tensor decomposition, which allows a different compression rate for each modality, enabling the model to adapt to changes in the amount of useful information between different modalities. With the same advantages as the tensor fusion method, the compression method reduces model complexity and decreases the number of parameters with minimal loss of information, so it can be used as a regularizer to avoid overfitting. Moreover, through the study of sensitivity to modality-specific compression rate, it is helpful in understanding the relative amount of non-redundant information in each modality.

#### 4.3.4. T2FN (Temporal Tensor Fusion Network)

Clean multimodal time series often show correlations across time and modalities, presenting a low-rank tensor representation [65]. However, the presence of imperfect modalities, missing entries, and noise corruption can break these natural correlations and lead to high-rank tensor representations. Therefore, Paul et al. [66] proposed a Temporal Tensor Fusion Network (T2FN) model based on a tensor rank minimization regularization method, which learns tensor representations of true correlations and latent structures in multimodal data and effectively normalizes their rank. T2FN extends the TFN model by adding a temporal component that enhances the ability to capture high-rank tensor representations, which in itself leads to improved prediction performance. The adaptation to imperfect data reflects the robustness of the model.

#### 4.3.5. HFFN (Hierarchical Feature Fusion Network)

Also inspired by TFN [21], in order to solve the problem brought by high-rank tensors, Mai et al. [23] proposed a Hierarchical Feature Fusion Network (HFFN) to improve efficiency through a hierarchical fusion framework. The whole model can be divided into three stages: “divide”, “conquer” and “combine”. The “divide” and “conquer” stages focus on local fusion, and the “combine” stage focuses on global fusion.

In the “divide” stage, the trimodal feature vectors are aligned to form a multimodal embedding and segmented into local parts using a sliding window to explore intermodal dynamics. In the “conquer” stage, the outer product is applied to fuse the features within each local block to explore the interactive state dynamics. Compared with other models that employ outer products, the efficiency is significantly improved due to the characteristics of the local model. In the “combine” stage, global interactions are modeled by exploring the interconnections and contextual dependencies between locally fused tensors. In response to the lack of connection between local fusion vectors, the authors proposed Attentive Bi-directional Skip-connected LSTM (ABS-LSTM), an RNN variant. ABS-LSTM introduces bidirectional skip connections of memory cells and hidden states into LSTM, and integrates an attention mechanism to transfer information more efficiently and learn holistic representations to obtain a holistic view of multimodal information. Finally, the global interaction is input to the emotional inference module to obtain the final prediction.

#### 4.3.6. LMFN (Locally Confined Modality Fusion Network)

In the second year of proposing HFFN [23], Mai et al. [24] presented a new multimodal fusion framework called the locally confined modality fusion network (LMFN). For intra-modal interactions, each modality has a specific unimodal context-dependent learning network UC-LSTM. Similar to BC-LSTM, UC-LSTM consists of a bidirectional LSTM layer and a dense layer, which can model the temporal relationship between consecutive utterances in a video to help better understand the current utterance. As in HFFN, the authors divided the fusion into local and global stages. In local fusion, the feature vector corresponding to each modality is divided into multiple segments, and each local interaction is learned through a tensor fusion process. The local vectors are then passed to the global fusion stage, where dependencies are learned by a Bidirectional Multiconnected LSTM (BM-LSTM) to model global interactions, establishing direct connections between cells and local tensor states that are several time steps apart. Finally, the output is connected to the decision layer to arrive at sentiment predictions.

### 4.4. Word-level fusion

The word-level fusion approach takes into account both view-specific and cross-view interactions, and efficiently explores time-dependent interactions by modeling interactions at each time step. The model framework of this fusion method generally consists of two modules: a temporal modeling module and an attention module.

In the temporal modeling module, modality-specific dynamics are modeled through a temporal modeling network (LSTM, LSTHM, 1D temporal CNN, etc.).

The attention module receives the output of the temporal modeling module and uses the attention mechanism and its variants to model important information in dynamic cross-modal interactions.

Table 6 summarizes the advantages and disadvantages of each model.

#### 4.4.1. GME-LSTM(A) (Gated Multimodal Embedding LSTM with Temporal Attention)

Diverging from previous multimodal sentiment analysis work that focused on the overall information of speech fragments, Chen et al. [25] proposed a Gated Multimodal Embedding LSTM with Temporal Attention (GME-LSTM(A)) model. This is the first method to perform multimodal fusion at the word level. The model is divided into two modules: the Gated Multimodal Embedding Layer and the LSTM with Temporal Attention module.

Gated multimodal embedding learns local interactions between modalities at each time step (word level). Since previous models have shown that noise in visual and acoustic modalities can impair the performance of the textual modality, the authors used an on/off input gate controller on the acoustic/visual features for selective multimodal fusion, which alleviates the difficulty of fusion in noisy modalities. Features from the three modalities are concatenated and then fed into the next layer. LSTM with temporal attention captures temporal interactions on multimodal embedding layers and adaptively focuses on the most important time steps. This module learns global interactions between modalities, enabling the model to account for both local and global interactions.

#### 4.4.2. MARN (Multi-Attention Recurrent Network)

Zadeh et al. [26] proposed a Multi-attention Recurrent Network (MARN), which consists of two key components: Long-short Term Hybrid Memory and Multi-attention Block. Long-short Term Hybrid Memory (LSTHM) is an extension of LSTM that redesigns its memory components so that mixed information can be carried. This hybrid memory can store not only intra-modal information for a specific modality but also important cross-view dynamics related to that modality. Among them, the component that discovers the interactive information of the transmembrane state is the Multi-attention Block (MAB), which is the uniqueness and advantage of this model. At time step t, the MAB accepts hidden states from all LSTHMs to outline the multiple cross-view dynamics that exist. The authors posit the K largest cross-view dynamics at each timestamp and use a deep neural network to obtain these K attention coefficients. Then, according to these coefficients, the output dimensions are weighted to form the dynamic code $z_t$ of the LSTHM at time t, which is passed to the next time step.

**Table 6**
Characteristics of word-level fusion methods.

| Methods | Advantages | Disadvantages |
|---|---|---|
| GME-LSTM(A) | • The first to use word level modality fusion<br>• Gated multimodal embedding filters out noise from acoustic and visual data<br>• LSTM with temporal attention performs word-level fusion at a finer fusion resolution between input modalities | • The temporal correlation of the individual modality is ignored<br>• Obtaining the word-level features needs to perform the force-alignment, which is time-consuming |
| MARN | • Long-short Term Hybrid Memory (LSTHM)<br>• Multi-attention Block (MAB)<br>• Discover interactions between modalities over time and store them in the hybrid memory of the recurrent component | • Cannot learn the correlations across different modalities |
| MFN | • Multi-view sequential learning<br>• Delta-memory Attention Network (DMAN)<br>• Memorize long-term interactions and internal behaviors across modalities, and store and update them in LSTM | • Cross-view dynamics are not explored |
| DFG | • Replace the original fusion component DMAN in MFN with DFG | • Useless visual modality can sometimes hurt model performance |
| RMFN | • Recurrent Multistage Fusion Network (RMFN)<br>• Decompose the fusion problem into multiple stages, each of which focuses on a subset of multimodal signals for specialized and efficient fusion<br>• The multistage fusion method can be easily extended to memory-based fusion methods | • Lack of representational power to accurately model the structure of nonverbal behavior at the subword level |
| HED | • A hierarchical encoder–decoder framework<br>• The encoder learns word-level features from each modality that are then formulated into conversation-level features<br>• The decoder decodes features to each time instance, further decomposing them into attributes for sentiment prediction | • Not suitable for small datasets<br>• Decoding the entire conversation into time instances with equal length results in the inability to make sentence-level predictions |
| RAVEN | • Nonverbal sub-networks<br>• Model the fine-grained structure of nonverbal subword sequences<br>• Use an attention gating mechanism to yield the nonverbal shift vectors which characterizes the extent and direction of word changed due to nonverbal context | • Only a simple LSTM is used for making predictions |
| SWRM | • Sentiment word position detection<br>• Apply multimodal gating network to filter out useless information from the input word embeddings<br>• Utilize useful information from candidate sentiment words as a supplement to the filtered word embeddings<br>• Can be adapted for other multimodal feature fusion models easily | • Only text is used to detect the position of sentiment words, no other modalities are exploited |

### 4.4.3. MFN (Memory Fusion Network)

Zadeh et al. [27] proposed a novel neural model for multi-view sequence learning, named the Memory Fusion Network (MFN). The sequential interactions of each modality over time are encoded using LSTMs. The output of the LSTM system is then concatenated into an attention layer to identify cross-view interactions by associating a relevance score with the memory dimension of each LSTM, called the Delta-memory Attention Network (DMAN). The outputs of the DMAN at time steps t-1 and t are passed to multi-view gated memory to indicate what dimensions in the memory system of LSTMs constitute cross-view interactions. The module then updates its content based on the output of the DMAN and its previously stored content, controlled by two sets of gates, both controlled by a neural network. At each time step, the retention gate assigns how much of the current state of the Multi-view Gated Memory to remember and the update gate assigns how much of the Multi-view Gated Memory to update based on the updated proposal. The output is the final state of the multi-view gated memory and the vector concatenation of the LSTMs representing the information of a single sequence.

### 4.4.4. DFG (Dynamic Fusion Graph)

In addition to proposing CMU-MOSEI, Zadeh et al. [9] defined the establishment of n-modal dynamics as a hierarchical process and proposed a new fusion model called the Dynamic Fusion Graph (DFG). The original fused DMAN in MFN [27] is replaced by DFG, and the final result model is called the Graph Memory Fusion Network (Graph-MFN).

### 4.4.5. RMFN (Recurrent Multistage Fusion Network)

Liang et al. [67] proposed a Recurrent Multistage Fusion Network (RMFN), which decomposes the multimodal fusion problem into multiple recursive stages, each focusing on a subset of multimodal signals for specialized and efficient fusion. Multistage fusion mainly consists of three modules: Highlight, Fuse, and Summarize. First, each modal sequence is modeled with an intra-modal recurrent neural network [26]. At time step t, each intra-modal recurrent network will output a unimodal representation. Then in the multistage fusion process, the two modules of Highlight and Fuse are repeated at each stage, where the Highlight module identifies subsets of multimodal signals, and the Fuse module fuses the highlighted features locally and combines them with the previous stage. Finally, the Summarize module converts the multimodal representation of the last stage into a transmembrane state representation, which is passed to the next time step. The final representation integrates the last output of the LSTHMs and the last transmembrane state representation. By combining this fusion method with a recurrent neural network system, interactions within time and patterns are simulated.

### 4.4.6. HED (Hierarchical Encoder-Decoder)

Gu et al. [68] proposed a hierarchical encoder–decoder structure with an attention mechanism for conversation analysis. The system consists of two modules: the hierarchical encoder and the hierarchical decoder. The hierarchical encoder first synchronizes and combines the extracted feature-shared representations, then uses a temporal attention mechanism to select important word vectors to form a single feature vector. In the hierarchical decoder, features are decoded to each temporal instance, and finally multi-label predictions are obtained.

### 4.4.7. RAVEN (Recurrent Attended Variation Embedding Network)

Considering that previous work [22,26,67] used a simple averaging strategy to summarize subword information during each word span and lacked the representational ability to accurately model the structure of nonverbal behaviors at the subword level, Wang et al. [28] proposed a Recurrent Attended Variation Embedding Network (RAVEN) which can be divided into three parts: Nonverbal Sub-networks, Gated Modality-mixing Network, and Multimodal Shifting. First, the non-verbal sub-network uses two independent LSTMs to encode the fine-grained structure of non-verbal behaviors, and generate embedding vectors. The gated modality-mixing network component then computes non-linguistic shift vectors by learning nonlinear combinations between visual and acoustic embeddings through an attentional gating mechanism. The significance of this vector is to describe the impact of non-linguistic context on text words. Finally, the non-linguistic shift vector is integrated into the original word embeddings to compute multimodal shifted word representations that dynamically capture contextual changes in different non-linguistic contexts. The final multimodal representation can be passed to a word-level LSTM for encoding and prediction through fully connected layers.

#### 4.4.8. SWRM (Sentiment Word Aware Multimodal Refinement Model)

In the real world, since textual content in videos is generally recognized by automatic speech recognition (ASR) models, some key emotional elements may be recognized as other words, resulting in a sharp drop in the performance of even advanced models. To address this problem, Wu et al. [29] developed three real-world datasets based on the existing dataset CMU-MOSI [8], using three widely used ASR APIs for text recognition, including SpeechBrain, IBM, and iFlytek. The authors also proposed a Sentiment Word Aware Multimodal Refinement Model (SWRM), which achieves good results on these datasets. The proposed model consists of three modules for sentiment word position detection, multimodal sentiment word refinement, and multimodal feature fusion respectively.

First, the sentiment word position detection module is used to obtain the most probable position of the sentiment words in the text. Since ASR may recognize sentiment words as neutral words, instead of sentiment word localization, the authors used the strong language model BERT to generate candidate sentiment words. Then, the multimodal sentiment word refinement module is used to refine sentiment word embeddings dynamically. The refinement process consists of two parts, filtering and adding. During the filtering process, a non-linear neural network called the multimodal gating network is applied to filter out the useless information in the input word embeddings. After the filtering process, a linear layer is used to extract useful information from candidate sentiment words, which are added to the word embeddings to produce a multimodal sentiment word attention network. Finally, the refined embedding data is used as the text input to the multimodal feature fusion module to predict sentiment labels.

### 4.5. Translation-based fusion

This category is a method of modeling the interaction between modalities employing translation. Inspired by the success of sequence-to-sequence (Seq2Seq) models in machine translation, researchers propose to convert one modality to another to capture more meaningful relationships across modalities. Another option is to use a pre-trained language model to capture word interactions by adjusting the structure of the transformer encoder. Table 7 summarizes the advantages and disadvantages of each model.

#### 4.5.1. MCTN (Multimodal Cyclic Translation Network)

Inspired by the unsupervised representation learning of the Seq2Seq model, Pham et al. [30] proposed a Multimodal Cyclic Translation Network (MCTN) model to learn a robust joint multimodal representation by transforming modalities. In addition to the forward transformations from source to target modality, the authors also added a backward transformation from predicted target to source modality to ensure that the learned joint representation can capture the maximum information from both modalities, which is named multimodal cyclic translations.

For the joint representation between the source modality and multiple target modalities, a hierarchical model is also proposed. In the first layer, a defined multimodal cyclic translation learning is employed to learn intermediate representations. After passing to the second layer, the backward transformation is canceled and only the forward transformation is used to obtain the final representation. Finally, the representation is fed into an RNN classifier to produce predictions. The advantage of MCTN is that once trained with multimodal data, only the data from the source modality needs to be used at test time to infer joint representations and labels. Therefore, MCTN is completely robust to test time perturbations or missing information on other modalities.

#### 4.5.2. MulT (Multimodal Transformer)

Tsai et al. [31] proposed a Multimodal Transformer (MulT), using directional pairwise crossmodal attention to realize the interactions between multimodal sequences across distinct time steps and latently adapt streams from one modality to another. A 1D temporal convolutional layer is first used to provide each element in the input sequence with sufficient knowledge of its neighbors. To carry time information, the authors also add position embedding (PE) to the sequence. The highlight of this paper is that a cross-modal transformer is proposed, which enables one modality to receive information from another modality by interacting with cross-modal attention. Each modality, interacts with the other two modalities, accepting low-level external information to continuously update its sequence. Finally, a self-attention transformer is used for cross-modal transformers with the same target modality to collect temporal information, concatenating the input to a fully connected layer for prediction.

#### 4.5.3. MAG (Multimodal Adaptation Gate)

With the superior performance of transformer-based contextual representations on downstream tasks, Rahman et al. [32] proposed a method to efficiently fine-tune large pre-trained transformer models for multimodal languages, termed Multimodal Adaptation Gate (MAG). MAG can be seen as an add-on to BERT and XLNet, allowing these to accept multimodal non-linguistic data during fine-tuning.

In the encoding layer, MAG accepts inputs from three modalities. The language vector is connected with acoustic and visual information respectively to form two bimodal factors. Furthermore, two gating vectors are generated, which embody the acoustic and visual information based on language. Multiplying the two gating vectors with their respective modality vectors yields a non-verbal displacement vector. In other words, MAG exploits the information of non-verbal behavior to form a vector with trajectory and amplitude to add to the verbal vector. This non-verbal displacement vector modifies the internal states of BERT and XLNet during fine-tuning of the pre-trained model, and finally obtains a multimodal vector. A unique feature of this method is that the MAG component is merely an add-on to BERT or XLNet without changing the original structure.

#### 4.5.4. TCSP (Text-Centered Shared-Private)

Departing from previous works that treat the features of the three modalities equally, Wu et al. [69] proposed a Text-centered Shared-private framework (TCSP). This model takes the text modality as the core and enhances the semantics of the text through the other two modalities. The authors divide the contributions of visual and acoustic modalities into shared semantics and private semantics. Shared semantics refers to the information common to the three modalities. Although it cannot provide additional information to text modalities, repeated information can significantly enhance text semantics. Private semantics are non-linguistic modality-specific semantics that can complement textual semantics to help detect the final sentiment more accurately. The framework consists of a cross-modal prediction task and a sentiment regression model.

Two cross-modal prediction models, namely text-to-visual and text-to-acoustic models, are trained to explore shared and private semantics from non-text modalities. Considering that shared semantics contain

**Table 7**
Characteristics of translation-based fusion methods.

| Methods | Advantages | Disadvantages |
|---|---|---|
| MCTN | • Multimodal cyclic translation<br>• The model learns increasingly discriminative joint representations with more input modalities while maintaining robustness to missing or perturbed modalities<br>• Only data from the source modality need to be used at test time for final sentiment prediction | • Cyclic translation between modalities requires a lot of training resources and time |
| MulT | • The cross-modal attention module fuses multimodal information by directly attending to low-level features in other modalities<br>• Can be directly applied to unaligned multimodal streams | • When part of the language information is missing, the model lacks continuous attention and generalization to the target modality, and the overall performance is slightly lower |
| MAG | • Non-verbal displacement vector<br>• MAG makes no change to the original structure of BERT or XLNet, but acts as an attachment to both models | • Both input-level concatenation and addition of modalities perform poorly |
| TCSP | • Treat the textual modality as the core and use the other non-textual modalities to help enrich the semantics of the textual modality<br>• Shared and private semantics<br>• Effectively fuse textual and non-textual features benefiting from unlabeled data | • The performance of cross-modal prediction model greatly affects the effectiveness of regression model |
| ICDN | • The encoder of the traditional Transformer is improved so that it can receive the input of multimodal information<br>• Modify the mapping attention to map the bimodal information to the target modality, so as to obtain richer modality-related information<br>• Improve self-supervised unimodal label generation module (ULGM) and uses unimodal for multi-task learning to supplement and generalize multimodal fusion | • After using Mapping Transformer, features containing rich information of the original modality are still required to further reduce modal differences |
| DEAN | • Deep emotional arousal network<br>• Model the sentiment congruence by introducing time-dependent interactions into the parallel structure of Transformer<br>• Identify the difference between different modalities by embedding a multimodal gating mechanism | • Time consumption of transformer-based model |
| AMOA | • The first to introduce the global acoustic features to enhance the learning of overall video features<br>• By designing a cross-modal transformer (CMT) that integrates the three modalities in a certain order, the modality added before can also provide information for the later processes<br>• Generalize to sentiment with more complex semantics, achieving SOTA performance on a widely used sarcasm dataset | • Need to know the importance of each modality in advance to determine the order of fusion |

more transmembrane state information, words assigned higher attentional weights are classified as shared. Since modal-private information is difficult to predict via text modality, features with higher prediction loss in non-linguistic modalities are considered private. After obtaining shared semantics and private semantics, a sentiment regression model is applied to fuse text features with those of the other two modalities. The sentiment regression model is mainly composed of three parts: shared module, private module, and regression layer. In the shared module, a masked cross-modal attention network is proposed to leverage the shared information of non-linguistic modalities to enhance the representation of words. Meanwhile, in the private module, the private features are passed through the attention layer to produce the final private representation. Finally, the obtained representation is input to the regression layer, which is a two-layer network with ReLU, to predict the sentiment score. It is worth mentioning that this work provides a new method for multimodal sentiment analysis using unlabeled data [69].

### 4.5.5. *ICDN (Integrating Consistency and Difference Networks)*

Inspired by the application of the Transformer [70] structure and Mult [31] in the multimodal domain, Zhang et al. [33] proposed a method known as Integrating Consistency and Difference Networks (ICDN). Instead of a sequence-to-sequence structure, the model consists of multiple mapping attention modules for deeper modality fusion based on the low-level features of each modality. Firstly, ICDN uses a Mapping Transformer (MT) to map the low-level features of the remaining two modalities to a third modality, making up for the loss caused by the missing parts of that modality. Unlike Mult, the MT module abandons the decoder and improves the encoder to obtain richer modality-related information using self-attention techniques. Second, Transformers are used to extract modal features, improving long-range dependencies between modalities and attention to contextual information. Finally, the self-supervised method in SELF-MM [58] is improved to obtain unimodal sentiment labels, and multi-task learning guides the final fusion of multimodal features.

### 4.5.6. *DEAN (Deep Emotional Arousal Network)*

Departing from most other models that focus on considering more efficient fusion strategies, inspired by the emotional arousal model in psychology, Zhang et al. [71] proposed a Deep Emotional Arousal Network (DEAN) to simulate the whole process of human communication at the multimodal input. DEAN consists of three components: a cross-modal transformer, a multimodal BiLSTM system, and a multimodal gating module, respectively simulating the functions of the perception analysis system, cognitive comparator, and activation mechanism in the psychological emotional arousal in humans. On the input side, the cross-modal transformer explores the interaction between every two modalities through an improved multi-head attention mechanism, that is, DEAN contains a total of 6 cross-modal transformers. Afterward, the authors adopted a multimodal BiLSTM system to extract the context-related features of each modality through a Bidirectional LSTM network and modeled the time series to simulate emotional coherence. The final module is the Multimodal Gating Block, which implicitly performs fusion between modalities by selecting the information to be output according to the importance of the modalities. DEAN attempts to provide a complete framework and an alternative thought that guides the learning system along a human-like path to progressively acquire a complex understanding of human emotions.

### 4.5.7. *AMOA (Acoustic feature en-hanced Modal-Order-Aware network)*

To cope with the problem that three modalities are equally treated by existing methods and the global acoustic information is lost after the video is divided into frames, Li et al. [34] proposed a global Acoustic feature en-hanced Modal-Order-Aware network (AMOA). After the features of each modality are extracted, a Transformer encoder-based Cross-Modal Transformer (CMT) is designed for two-stage fusion. In

**Table 8**
Characteristics of feature space manipulation-based fusion methods.

| Methods | Advantages | Disadvantages |
|---|---|---|
| ICCN | • Learn correlations between all three modalities via Deep Canonical Correlation Analysis (DCCA)<br>• Based on the text, two outer-product matrices are built for representing the interactions between text-video and between text-audio | • The dynamic intra-actions in each model and the inter-actions between different modalities are for further study<br>• There needs to be a trade-off between maximum canonical correlation and optimal downstream task performance |
| MISA | • Each modality is decomposed into modality-invariant and modality-specific features, so as to learn its common and private attributes respectively<br>• Stress the importance of representation learning before fusion<br>• Effort made in exploring the feature space reduce the need for complex fusion mechanisms | • Not much contribution in exploring more efficient fusion solutions |
| HyCon | • The first to leverage contrastive learning in a hybrid manner to learn cross-modal embeddings<br>• Pair selection mechanism<br>• Inter-class and inter-sample relationships are explored for a more discriminative joint embedding<br>• The designed loss function introduce no additional parameters, which reduces the possibility of overfitting and improves the generalization ability | • Lack of exploring potential relationships within and between modalities<br>• The model takes slightly longer to run than most baselines<br>• Little attention has been paid to the design of new fusion methods |
| HGraph-CL | • Hierarchical Graph Contrastive Learning (HGraph-CL) framework<br>• The construction of intra- and inter-modality graphs is explored to exploit potential sentiment relationships within and across modalities<br>• Better generalizability, transferability, and robustness in learning sentiment cues compared with pure class-driven methods | • Experimental results are sensitive to deleting/adding ratio |
| HIS-MSA | • Adopt different self-supervised pre-training strategies to fully mine the unique knowledge of the in-domain corpus<br>• A unimodal label generation module is used to jointly guide multimodal tasks and unimodal tasks to balance independent and complementary information between the modalities<br>• Heterogeneous graphs are introduced in the modality fusion phase to efficiently fuse information from multiple modalities | • Rely heavily on the quality of the parser<br>• There exists noise in the visual and audio features learned by the model |

the first stage, the text modality is the core, the acoustic modality is integrated, and then the visual modality is added in the second stage. One of the benefits of using CMT is that previously added modalities can provide information for later. In this way, the textual features are continuously enhanced while also reducing the noise in the visual modality. Since the resulting multimodal features are composed of the features of a single frame and lack overall acoustic information to reflect the overall change in tone, the authors add a Global Acoustic Feature (GAF). The GAF extracted with open SMILE is in a different space from the multimodal fusion feature, so the authors use contrastive learning to align the two features and concatenate them. Finally, the contrastive loss and the classification loss are added together to guide the model training, and a multilayer perceptron layer is used for classification.

## 4.6. *Feature space manipulation-based fusion*

This type of fusion method focuses on mapping features into feature space after feature extraction and learning the relationship between features through a series of mathematical analyses or operations. Table 8 summarizes the advantages and disadvantages of each model.

### 4.6.1. *ICCN (Interaction Canonical Correlation Network)*

Sun et al. [35] proposed an Interaction Canonical Correlation Network (ICCN) model to learn the correlations between all three modalities through Deep Canonical Correlation Analysis (DCCA). Canonical Correlation Analysis (CCA) is a well-known method to find the linear subspace with the largest correlation between two inputs [72]. On its basis, DCCA uses a pair of neural networks to learn nonlinear transformations, which solves the limitation of considering only linear transformations. After extracting audio and video through 1D convolution and LSTM, the outer product operation is performed with the text embedding. The resulting two representations are fed into a DCCA consisting of two CNNs and CAA layers to learn useful features in the outer product matrix. Finally, the output of DCCA is concatenated with the original text sentence embedding as the final multimodal embedding for sentiment classification.

### 4.6.2. *MISA (Modality-Invariant and -Specific Subspaces)*

Complex fusion methods are easily challenged by morphological gaps between different modalities. To solve this problem, Hazarika et al. [36] proposed a new framework, MISA, which considers different modal subspaces to improve the fusion effect. The framework can be divided into modal representation learning and modal fusion, where the former is the main contribution of the paper. After extracting the features of the three modalities, each modality is projected to two different subspaces. The first subspace is modality-invariant, and the distribution similarity constraint is applied to minimize the heterogeneity gap and learn their commonality. The second subspace is modality-specific and learns feature information private to each modality. After projecting the modalities into the corresponding subspace, a transformer-based self-attention is used to concatenate all 6 transformed mode vectors to make predictions. Despite the inclusion of simple feedforward layers, MISA's efforts to explore the feature space reduce the need for complex fusion mechanisms.

### 4.6.3. *HyCon (Hybrid Contrastive Learning)*

Trapped by the gaps in cross-modal information, most previous work has focused on exploring the interactions within and between modalities, while ignoring the learning of inter-sample and inter-class relationships. To address these issues, Mai et al. [37] proposed a new framework, HyCon, to learn trimodal representations using hybrid contrasts. After obtaining unimodal representations for each modality, the authors used three different contrastive learning models to learn inter-modality interactions and inter-class relationships. Semi-Contrastive Learning (SCL) only considers positive samples and learns the interactions between different modalities of the same sample in an unsupervised form. Intra-modality Contrastive Learning (IAMCL) and Inter-modal Contrastive Learning (IEMCL) both operate in a supervised manner; the former learns the intra-modality dynamics among different samples, and the latter learns the inter-modality dynamics. Both IAMCL and IEMCL explore inter-class relationships. Finally, various fusion strategies (including simple ones such as concatenation) are adopted to verify the effectiveness and generalizability of the model.

The authors also introduced refinement terms, modality margins, and pairing selection mechanisms to enhance the performance of the entire system; details on these additions can be found in the article [37].

#### 4.6.4. HGraph-CL (Hierarchical Graph Contrastive Learning)

Lin et al. [38] believe that existing work mainly focuses on fusing different modal information through class-driven supervised learning or multi-task learning, but this approach cannot understand complex relationships within and across modalities. They proposed a novel Hierarchical Graph Contrastive Learning (HGraph-CL) framework to address this difficulty, where highly correlated modal representations are explicitly linked. For intra-modal dynamics, a unimodal graph is constructed for each modality, where the text modality graph uses the grammatical dependency tree of the sentence, and the other two modalities exploit their continuous sequential relationships. Thus, combined with three modality-specific views, an inter-modal view is constructed for each multimodal instance by full connection to capture the potential dispersion of emotions. The authors then employ a graph attention network to model semantic relationships by assigning different weights to different nodes in the neighborhood. Following that, two forms of supervised contrastive learning strategies are applied to the intra- and inter-modal levels. One is to utilize sentiment labels as supervision signals and perform a fully supervised loss to capture the similarity of examples within a class as well as the contrast between classes. The other is a self-supervised graph contrastive learning strategy based on graph augmentation, which explores a more suitable graph structure by adding or removing edges. This hierarchical graph contrastive learning strategy enhances the learning of graph representations at both the data level and label level. Compared with pure class-driven methods, it has better generalization, transferability, and robustness in learning sentiment cues. More details can be found in the paper if the reader is interested.

#### 4.6.5. HIS-MSA (Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis)

The inability to make full use of domain knowledge and the lack of effective integration methods have been the difficulties and key points of multimodal sentiment analysis, so to solve this problem, Zeng et al. [73] proposed a heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis (HIS-MSA). First of all, to make the feature encoder better used for embedding text modality, based on the original BERT, professional knowledge is added, and different self-supervised training strategies are used for the second pre-training to obtain domain awareness. In the fusion stage, the authors take the syntactic dependency tree of text modality as the core to construct a heterogeneous graph, to integrate the features of the other two modalities. Heterogeneous graph convolutional networks can learn complementary information among multiple modalities interactively by driving heterogeneous graphs. Finally, inspired by previous work, multimodal tasks and unimodal tasks are jointly guided by a unimodal label generation module to balance independent and complementary information between modalities.

### 4.7. Contextual-based fusion

Previous methods treat each utterance as an independent entity and ignore the dependencies between utterances in the video. Contextual-based fusion achieves better results by considering the connections between other utterances in the context and the target utterance. Among them, recurrent neural network-based models are generally used to incorporate contextual information. Table 9 summarizes the advantages and disadvantages of each model.

#### 4.7.1. BC-LSTM (Bi-directional Contextual LSTM)

Given some works that treat utterances as independent entities and ignore the interrelations between utterances, Poria et al. [39] proposed a BC-LSTM model to capture the contextual information of utterances in the same video environment. Unimodal features that do not contain contextual information were first extracted by different feature extractors, then input into an LSTM network. The authors replaced the regular LSTM with a bi-directional LSTM so that an utterance could gain information from preceding and following utterances, naming their system bi-directional contextual LSTM (BC-LSTM) for this reason. Finally, the obtained unimodal features containing contextual information were concatenated and passed to a similar independent BC-LSTM for training, and sentiment classification results were output.

#### 4.7.2. CHFusion (Context-aware Hierarchical Fusion)

Majumder et al. [74] proposed a Context-aware Hierarchical Fusion (CHFusion) model, which uses a hierarchical structure to continuously fuse multi-modality information and update the context information after each layer fusion. First, the unimodal features of each utterance of the three modalities are obtained. Then, a GRU is used to extract context-aware discourse features. By combining the unimodal features containing contextual information in pairs through a fully connected layer, a bimodal feature vector is formed after fusion. As in the unimodal case, the GRU is also used to sense context. Finally, three bimodal fusion vectors are combined into a trimodal vector through a fully connected layer, and a GRU is used to convey contextual information. The output of the model is generated from a softmax layer.

#### 4.7.3. MMMU-BA (Multi-modal Multi-utterance-Bi-Modal Attention)

Ghosal et al. [40] proposed a Multi-Modal Multi-Utterance-Bi-Modal Attention (MMMU-BA) framework to leverage contextual information for utterance-level sentiment prediction. Unlike previous methods that simply apply attention to contextual utterances for classification, the authors focus on contextual utterances by computing the correlation between target utterance patterns and contextual utterances. After inputting the continuous utterances of three modalities into three independent bi-directional gated recurrent units (Bi-GRU [75]) and passing through a dense layer, three matrices containing the contextual information of specific modal utterances are obtained. Multimodal attention is applied to the output matrix of the dense layer to learn the multimodal connections between utterances. The "bi-modal attention" aspect of the framework lies in its application of an attention mechanism to paired modal representations, resulting in three sets of interactions, namely visual-text, text-acoustic, and acoustic-visual. Finally, the pairwise modal attention and the unimodal representation are concatenated and passed to the softmax layer for sentiment classification.

#### 4.7.4. CIA (Context-aware Interactive Attention)

Chauhan et al. [76] proposed an end-to-end Context-aware Interactive Attention (CIA) based recurrent neural network for sentiment analysis. The authors believe that the encoded representation between different modalities can learn their interactions, so the interaction between modalities is learned through an autoencoder-like structure called the inter-modal interactive module (IIM). IIM encodes a feature representation for one modality (e.g., text) and aims to decode it into a feature representation of another modality (e.g., vision). Thus, a total of 6 modal pairs are generated. The text-acoustic pairs and the acoustic-text pairs differ because the encoder contains two error gradients, one from the output of IIM $l_1$, and the other from the task-specific label $l_2$. Afterward, the sequential patterns of utterances are extracted by a bi-directional gated recurrent unit (Bi-GRU), and for each pair of modality interactions (e.g., text-acoustic and acoustic-text), an averaging operation is used to reduce the presence of dimension.

In addition, the raw data of the three modalities are also processed by a separate Bi-GRU, and the outputs are paired and then transmitted to a fully connected layer to extract the Bimodal Interaction (BI). The

**Table 9**
Characteristics of contextual-based fusion methods.

| Methods | Advantages | Disadvantages |
|---|---|---|
| BC-LSTM | • A LSTM-based framework is developed to extract contextual utterance-level features<br>• The model preserves the sequential order of utterances and enables consecutive utterances to share information | • The importance of each utterance and its specific contribution to each modality is not considered |
| CHFusion | • The hierarchical fusion structure makes every pair of modalities interact and combine into a trimodal vector, which captures the interrelationship between modalities<br>• Each layer uses RNN to extract context-aware utterance features | • The quality of unimodal features could be improved<br>• Simple network architecture |
| MMMU-BA | • The first work that attempts to employ multi-modal attention block (exploiting neighboring utterances) for sentiment prediction<br>• Focus on contextual utterances by computing correlations among the modalities of the target utterance and the context utterances, thereby helping to distinguish which modalities of relevant contextual utterances are more important<br>• Apply attention to multi-modal multi-utterance representations in an attempt to learn the contributing features | • When experimenting on the MOSEI dataset, the performance of the negative category is poor |
| CIA | • An end-to-end Context-aware Interactive Attention (CIA) based recurrent neural network that identifies and assigns the weights to the neighboring utterances based on their contributing features<br>• Learn the inter-modal interaction among the participating modalities through an auto-encoder mechanism<br>• Two affect analysis tasks were performed on five standard multi-modal affect analysis datasets | • Current work only applies to single-party utterances |

**Table 10**
Characteristics of quantum-based fusion methods.

| Methods | Advantages | Drawbacks |
|---|---|---|
| QMSA | • The first to apply Quantum Theory (QT) to sentiment analysis<br>• Images and text are encapsulated into density matrices, which can encode more semantic information and fill the “semantic gap”<br>• The multimodal decision fusion process is analogous to a double-slit experiment, and a Quantum Interference inspired Multimodal Decision Fusion (QIMF) strategy is proposed | • The computation time used for training and classification is longer than the use of other baselines<br>• The change of $\cos\theta$ has a great influence on the experimental results |
| QMN | • Quantum probability theory in the LSTM architecture is used to model both intra- and inter-utterance interaction dynamics<br>• A quantum measurement-inspired strong–weak influence model is proposed to make better inferences about social influence among speakers<br>• A quantum interference-inspired multimodal decision fusion method is proposed to model decision correlations between different modalities | • The model is largely dependent on the density matrix representation, how to further accurately capture the interactions among speakers and naturally incorporate them into an end-to-end framework is difficult<br>• Experiments are performed on emotion recognition datasets, not sentiment analysis datasets |
| QMF | • The interaction within a single modality and the interaction across modalities are formulated with superposition and entanglement respectively at different stages<br>• Sentiment decisions are made via the concept of quantum measurement | • The quality of the extracted visual and acoustic features is not high<br>• Inconsistencies with quantum theory |

two representations of each pair of modalities are transferred to a context-aware attention module (CAM) to extract the correspondences of adjacent utterances. The attention module helps the network focus on the contribution features by weighting the current and adjacent utterances in the video. Finally, the output of CAM is concatenated and passed to the output layer for prediction.

### *4.8. Quantum-based fusion*

Existing methods are mainly based on neural networks that model multimodal interactions implicitly and incomprehensibly. Neural architectures allow models to learn multimodal interactions from large-scale data in an end-to-end manner, often resulting in satisfactory accuracy. However, multimodal interactions are implicitly encoded by these models, working like a black box with few numerical constraints, which increases the difficulty of understanding multimodal interactions in human language. As these models bring significant performance gains, researchers are looking for ways to understand the model to know if we can trust it and deploy it in real works [77], or if it includes privacy or security issues [78]. Thus, they began to study quantum-based multimodal fusion methods. Table 10 summarizes the advantages and disadvantages of each model.

#### *4.8.1. QMSA (Quantum-inspired Multimodal Sentiment Analysis)*

Zhang et al. [79] proposed a Quantum-inspired Multimodal Sentiment Analysis (QMSA) framework, which is the first work to apply Quantum Theory (QT) to sentiment analysis. The framework consists of a Quantum-inspired Multimodal Representation (QMR) model and a Quantum Interference-inspired Multimodal Decision Fusion (QIMF) strategy. In the first part, the QMR model, based on the Quantum Language Model (QLM), represents texts and images through respective density matrices. For images, pixels are extracted to construct visual words, which are packed into a density matrix after vector space mapping. This density matrix describes the probability distribution of visual words in the image. A similar approach is used for text. Compared with traditional vector-based representation models, QMR models can encode more semantic relations. In the second part, the authors apply QIMF to fuse decision-making. The QIMF strategy is inspired by the double-slit experiment, where the sentiment label of multimodal data is analogous to a photon. The sentiment of text and images is seen as two slits, and each sentiment score is a location on the detection screen.

#### *4.8.2. QMN (Quantum-like Multimodal Networks)*

Building on previous work, Zhang et al. [80] proposed a new framework for multimodal sentiment analysis, termed quantum-like multimodal network (QMN), leveraging the formalism of quantum theory and the LSTM architecture. First, based on QMSA [79], QMN develops a density matrix-based convolutional neural network (DM-CNN) to represent the text and images of all utterances in a video and serve as the input of the whole model. Second, inspired by quantum measurement theory, a strong–weak influence model is introduced. This structure measures the influence relationship between speakers and acts as a complement to the output gate of the LSTM unit. Thus, textual and visual features can be input into two LSTMs respectively, and local sentiment analysis results can be obtained. Finally, a QIMF method is designed to derive the final decision based on local results. This part is similar to the fusion decision in QMSA.

#### 4.8.3. QMF(Quantum-inspired Multimodal Fusion)

Li et al. [81] proposed a fundamentally new framework to address the shortcomings of neural networks inspired by quantum theory, which incorporates a principled approach to modeling complex interactions and correlations. In this quantum-inspired framework, the mode-specific dynamics and interactions between different modes are represented by superposition and entanglement at different stages, respectively. This framework explains advancing the understanding of multimodal interactions from both a quantum and a classical perspective.

### 4.9. Summary of different fusion methods

Early multimodal sentiment analysis methods are mainly divided into early fusion and late fusion. These two types of fusion are relatively simple, without a very complicated fusion framework. Early fusion is also called feature-level fusion. At the input end, the feature vectors of the three modalities are spliced together as the input features of the entire model. The feature is transmitted to a subsequent classifier for sentiment classification, which can be an SVM or some other deep learning network. The benefit of this form of fusion is that it only needs to consider how to design classifiers more efficiently, without any specific model design. However, there is an obvious drawback, that is, premature fusion of features from different modalities leads to a lack of detailed modeling of specific view dynamics, which in turn affects the modeling of cross-view dynamics and leads to overfitting. The process of late fusion can be said to be just the opposite. First, sentiment predictions are made for each modality, and then the results based on these predictions are integrated into the final result through different decision-making methods, so it can also be called decision-level fusion. These decision methods can be average, majority, weighted, or other statistical strategies. Thus, this form of fusion is strong in modeling view-specific dynamics. Thanks to the integration of its modules, it can adapt well to changes in the number of modalities. However, the result is that dynamic interactions across views cannot be fully explored, and low-level interactions between different modalities are ignored.

Tensor-based methods take advantage of tensor representation and interaction. After feature representation for each modality, the tensor product is computed. Tensors can capture important higher-order interactions across time, feature dimensions, and multiple modalities during the mapping process, and have strong capabilities in exploring cross-modal dynamics. However, the tensor-based fusion method requires a lot of computing resources to calculate the outer dot product, resulting in the exponential growth of computational complexity. There is also no fine-grained word-level interaction during fusion. Therefore, methods under this framework mainly focus on reducing the computational complexity and resource consumption of fusion to enable better generalization.

Since the above three types of fusion methods lack finer-grained interactions, we classify them into the category of utterance-level fusion methods. The three categories of methods summarized next pay more attention to the fine-grained interactions of modalities, which we describe as fine-grained fusion methods, including word-level fusion methods, translation-based fusion methods, and FSM-based fusion methods.

Word-level fusion method models the interaction relationship at each time step and extracts useful information through an attention mechanism. Therefore, the framework generally consists of a temporal modeling module and an attention module. The temporal modeling module includes temporal networks such as LSTM and 1D temporal CNN. This module focuses on exploring modality-specific dynamics. The attention module employs the attention mechanism and its variants to model important information between modalities to mine cross-modal interactions. Word-level fusion methods effectively explore time-dependent interactions, but combining unimodal features with timestamps would miss an explicit and separate component to handle intra-modal and inter-modal interactions.

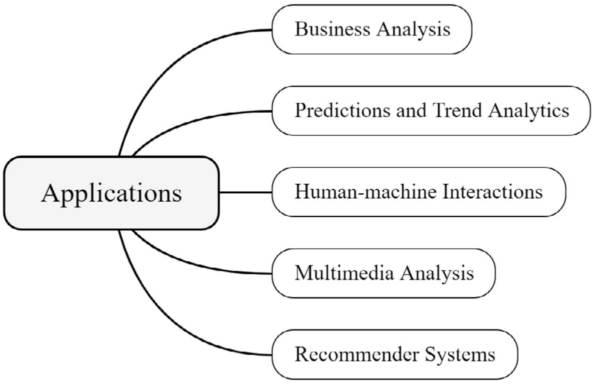

**Fig. 3.** Applications of multimodal sentiment analysis.

Translation-based fusion method is inspired by the Seq2Seq model in the field of machine translation. By translating one modality to another, more meaningful relationships across modalities are mined. The transformation process can complement the missing information of the modality, thereby enhancing the meaning. However, the translation based on the correlation of individual word representations will largely ignore the word order information. Another form is to capture word interactions by adjusting the structure of the pre-trained language model or adding additional components.

The full name of the FSM-based fusion is Feature Space Manipulation-based fusion. Through a series of mathematical analyses or learning models, it focuses on exploring the relationship between features in the feature space. The advantage of this fusion method is that it has a strong ability to explore the interaction between features, but the model does not consider an effective fusion strategy.

Contextual-based fusion can also be called multi-utterance fusion, because it not only considers the target utterance but also combines other utterances in the context. Generally, models based on recurrent neural networks are used to focus on contextual information. The target utterance and other utterances in the context form a context sequence, which can help judge the polarity of the target utterance more effectively. However, insufficient consideration is given to utterance-level sentiment analysis, and the extraction of contextual relations can easily lead to overfitting.

Quantum-based fusion approaches differ from existing neural networks in modeling multimodal interactions implicitly and incomprehensibly. Modeling multimodal interactions in quantum-inspired ways such as superposition, entanglement, and interference can resolve the paradoxes of classical probability theory in modeling human cognition while having better interpretability. However, there are many paradoxes in quantum theory, and the analogies in sentiment analysis do not quite line up.

## 5. Applications of multimodal sentiment analysis

Sentiment analysis, including multimodal sentiment analysis, is commonly used in business to summarize customers' opinions about a product or brand, as is multimodal sentiment analysis. Using automated sentiment analysis, we can obtain feedback from customers in a low-cost way. Previously, early-stage companies or organizations typically assessed customer opinions through survey panels, a relatively tedious and costly task. With the development of social media and the trend of people keen to post opinion videos and comments on, e.g., YouTube, automatic sentiment analysis can become a low-cost job. Multimodal sentiment analysis has also derived a series of applications and fields [82–84]. Following are some applications of multimodal sentiment analysis as shown in Fig. 3.

### 5.1. Business analysis

Multimodal sentiment analysis has many applications in the field of business intelligence. The most typical application is to analyze customer evaluations of products or brands. However, these studies are not only available to product producers; consumers can also use them to judge the quality of pre-order items and make more informed decisions [85]. For example, companies can leverage multimodal sentiment analysis data to improve products, investigate customer feedback and develop innovative marketing strategies [86]. Multimodal sentiment analysis can help customers choose better products by defining keywords for specific topics and training a sentiment analysis framework that can identify and analyze only the necessary information [87]. In addition, multimodal sentiment analysis can be applied to judge potential competitors and compare marketing methods, and obtain information on major consumers through user portraits [88] and other methods.

### 5.2. Predictions and trend analytics

Tracking public opinion through sentiment analysis can help predict some market scenarios. For example, analyzing video reviews of movies provides an opportunity to predict the box office performance of movies. In [89], the authors used Weka's KMeans clustering tool on Twitter, YouTube, and the IMDB movie database to generate movie box office predictions. The volatility and uncertainty of the stock market make stock market forecasting a daunting task, but by analyzing all the news about the stock market, the overall polarity of a particular company can be determined, thereby predicting stock price trends. Xing et al. [90] proposed such an analysis in their survey, associating positive news with an upward trend and negative news with a downward trend. Ma et al. [91] proposed a novel Multi-source Aggregated Classification (MAC) method to predict stock price movements by combining numerical characteristics and market-driven news sentiment of target stocks, as well as news sentiment of related stocks.

### 5.3. Human–machine interactions

The work of Langlet et al. [92] suggests that humans are likely to find an Embodied Conversational Agent (ECA) more likable if the user and the ECA share a common view of an entity. The field of human–machine interaction is also a large area of applied sentiment analysis. During the interaction between the avatars represented by ECA and the user, the management of the sentimental component of the dialogue is crucial. For different users, sentiment analysis can extract expressions of like or dislike. In this way, the trained ECA can better fit the user's psychology and provide better service effects. ECAs have found their way into many different applications, from online education to customer service.

### 5.4. Multimedia analysis

Multimedia analysis is a new field of development for sentiment analysis. The work of Ellis et al. [93] builds a system using multimodal sentiment analysis that can automatically analyze broadcast video news and create summaries of TV programs. Multimodal sentiment analysis techniques can also be used to identify politically persuasive content. Siddiquie et al. [94] proposed a solution to detect politically persuasive videos posted on social media and develop methods to predict and analyze comment reaction sentiment. The existence of multimodal sentiment analysis makes it possible to mine the opinions expressed by countless broadcast TV channels or online channels on the Internet.

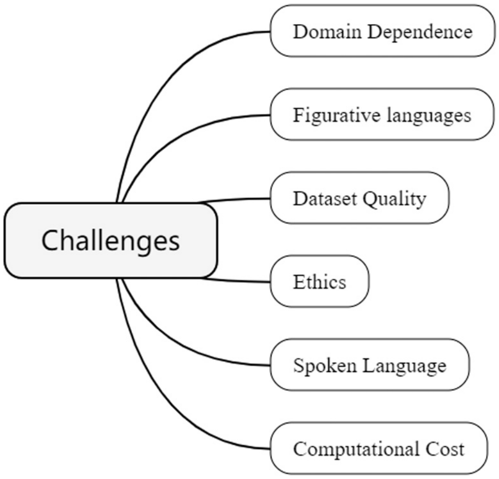

**Fig. 4.** Challenges of multimodal sentiment analysis.

### 5.5. Recommender systems

Many applications will give corresponding recommendations based on the user's historical search experience. For example, Amazon uses recommendation systems to recommend products on the homepage for customers, YouTube recommends related videos to play next on autoplay, and Facebook recommends people and web pages of potential interest. If a user searches for a specific product, that app will be suggested in future results. Dang et al. [95] proposed that incorporating sentiment analysis into recommender systems can significantly improve recommendation quality, especially when only sparse data are available.

## 6. Challenges

Combining information from different modalities is a challenging task, and we need to decide which modality holds more weight. It is also very important to reduce noise data between heterogeneous input data, which necessitates the design of better fusion methods and models. Apart from this, there are some other challenges in the field of multimodal sentiment analysis, which are introduced in this section and shown in Fig. 4.

### 6.1. Domain dependence

Most of the current sentiment analysis models are data-driven, so the model only learns the knowledge contained in the specific domain data. When a model trained in a specific domain is transferred to other unrelated domains, its performance will often drop significantly. Solving the domain-dependent problem requires designing a domain-dependence sentiment analysis system, for example, adapting a model trained on sentiment analysis in product reviews to analyze Weibo posts. Researchers have proposed many possible solutions, among which the prompt-based approach is a promising research direction. Pre-trained language models (PLMs) are trained with non-domain-specific texts, and adding prompts can help adapt to different domains. Mao et al. conducted a systematic empirical study of prompt-based sentiment analysis and emotion detection to investigate the bias of PLMs to affective computing [96].

### 6.2. Figurative languages

Analyzing figurative language including ambiguity, irony, and metaphor remains an inaccurate task in the field of multimodal sentiment analysis. Ambiguity and irony pose a major challenge to sentiment analysis. For example, comments ostensibly praising an object may be intended to convey a negative emotion; however, traditional sentiment analysis methods often misinterpret these expressions, judging them as positive. Many methods have been proposed to detect irony in language [97–99]. However, this problem is far from resolved, as many factors can affect irony, such as tone, situation, background information, etc., and humor is so culturally specific that it is challenging for machines to understand unique (and often very specific) cultural allusions. In the study of Poria et al. [100], it was proposed that by incorporating voice and facial expressions into multimodal sentiment analysis, its success rate in recognizing ironic comments could be improved. In addition, machine-based sentiment analysis results can only be based on external expressions of emotions, but cannot determine summative information about the thoughts expressed by an individual. Failure to recognize metaphors in sentences can also have disastrous effects on the accuracy of sentiment analysis. Because metaphors are often expressed differently from the conventional meaning of words, it may lead to the opposite state of the machine's judgment of sentiment. In the article [101], Mao et al. proposed a metaphor processing model called MetaPro, which can identify metaphors in sentences at the token level, paraphrasing the identified metaphors into their literal counterparts, and explain metaphorical multi-word expressions.

### 6.3. Dataset quality

A prominent data source for multimodal sentiment analysis is multimedia content on social media. Social media is a rich data repository that provides us with a sizable amount of data. However, the recorded material varies in quality and context, and the data is limited to statistics on specific groups of people on the Internet. Since this data is publicly available, it is easy to crowdsource tagging. Another source of data is private data from lab records, but this limits the tedious task of labeling to those authorized to access the data. Thus, in addition to limitations on the amount of data collected, the ability to tag large amounts of data is also limited. Grosman et al. proposed a new web-based text annotation tool ERAS. In addition to realizing the main functions of mainstream annotation systems, it also integrates a series of mechanisms to improve the annotation process and the quality of the annotation dataset itself, such as random document selection, Re-annotate stages, and warm-up annotations [102].

### 6.4. Ethics

Sentiment is a private state, and mining people's private states can raise ethical questions. Machines have incredible potential for understanding people's opinions and attitudes, but their use also raises questions about privacy. Sentiment analysis, as a data-driven technique, may introduce bias in decisions or higher-level analysis [7]. For example, if more Black men expressed strong opinions on Twitter, companies might pay more attention to their attitudes, since machine learning tools often treat data agnostically [103]. In addition, automated sentiment analysis could also be a tool to limit freedom of speech. What people say on social networks about policy opposition, partisan choices, etc. could be identified by sentiment analysis and censored at scale by an oppressive regime [104]. Other ethical issues that accompany data acquisition and annotation, such as evaluation in the development and long-term use of real-life recognition engines, are rarely explored. An experiment in [105] illustrates that the decision on the material used to stimulate or induce emotion may be critical, as certain materials or ways of eliciting responses may not be suitable for all participants in a database collection. For example, showing participants extreme violence may have a strong emotion-inducing effect, but may not be appropriate in many situations.

### 6.5. Spoken language

Another challenge of multimodal sentiment analysis is to effectively explore the intramodal dynamics of a specific modality. Since multimodal sentiment analysis is performed on spoken language, intramodal dynamic analysis of language is particularly challenging. A verbal opinion such as "I think it was alright . . . Hmmm . . . let me think . . . yeah . . . no . . . ok yeah" rarely happens in the written word. That is to say, this kind of the unstable verbal point of view is different from the rigorous written expression with good grammatical structure, but will be mixed with many habits of individual oral expression statements, and some meaningless modal particles will lead to proper language Structure is often overlooked, complicating sentiment analysis. Zhang et al. proposed to employ a deep reinforcement learning mechanism to select effective sentiment-relevant words and completely remove invalid words for each modality [106].

### 6.6. Computational cost

In order to attain higher accuracy and better results, we need to increase the size of the dataset and design a more efficient model [85]. However, the complexity of the model will lead to an exponential increase in the computational cost of training it. On the one hand, high-end GPU equipment is required to train a model with a huge corpus. On the other hand, high-complexity models have difficulty matching some specific scenarios. Traditional models such as SVM and NB are not computationally expensive, but the corresponding results are not good. Conversely, today's popular neural networks and attention models are computationally expensive. As a result, researchers will need to devise new models to balance performance and efficiency. For example, Han et al. proposed a novel encoder combining a hierarchical attention mechanism and feed-forward neural network to detect depressed individuals, which uses fewer training parameters than classical encoders [107]. Arjmand et al. proposed a transformer-based speech-prefixed language model with a lightweight attentive aggregation module to generate efficient spatial encoding. The model can achieve the same level of performance as taking a long time to retrain the Transformer without training a full Transformer model [108].

## 7. Conclusion and future work

This paper reviews recent advances in the field of multimodal sentiment analysis. We discuss the most popular datasets and feature extraction methods in the field. Recently published and cited articles are categorized and summarized according to the fusion method. The article further discusses available applications and the challenges faced by existing methods. Our review of the existing literature shows that multimodal sentiment analysis is a promising approach to leverage complementary information channels for sentiment analysis and often outperforms unimodal approaches. It also has the potential to enhance other tools that currently benefit from unimodal sentiment analysis, such as entity recognition and subjectivity analysis. We hope that this review will encourage further interdisciplinary efforts in this area.

An area worth exploring for future work is understanding sentiment in conversations. In a conversation, the sentiment one person expresses affects the others. Related work has demonstrated that discourse context is helpful for understanding human language, and if multimodal systems can simulate human sentimental dependencies, this will lead to significant progress in multimodal sentiment research. Further work is also needed to focus on making models language-independent to be able to generalize to any language in prediction tasks.

## CRediT authorship contribution statement

**Linan Zhu:** Conceptualization, Writing – original draft, Methodology, Investigation, Resources. **Zhechao Zhu:** Methodology, Writing – original draft, Writing – review & editing, Visualization. **Chenwei Zhang:** Formal analysis, Validation, Visualization. **Yifei Xu:** Validation, Visualization. **Xiangjie Kong:** Supervision, Project administration, Funding acquisition, Resources.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

No data was used for the research described in the article.

## Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (No. 62176234, 62072409).

## References

[1] L. Zhu, M. Xu, Y. Bao, Y. Xu, X. Kong, Deep learning for aspect-based sentiment analysis: a review, PeerJ Comput. Sci. 8 (2022) e1044.

[2] S.K. D'Mello, J.M. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Comput. Surv. 47 (3) (2015) 43:1–43:36.

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

[4] L. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: harvesting opinions from the web, in: Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, 2011, ACM, 2011, pp. 169–176.

[5] J. Yuan, M. Liberman, et al., Speaker identification on the SCOTUS corpus, J. Acoust. Soc. Am. 123 (5) (2008) 3878.

[6] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.

[7] M. Soleymani, D. García, B. Jou, B.W. Schuller, S. Chang, M. Pantic, A survey of multimodal sentiment analysis, Image Vis. Comput. 65 (2017) 3–14.

[8] A. Zadeh, R. Zellers, E. Pincus, L. Morency, MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, CoRR abs/1606.06259.

[9] A. Zadeh, P.P. Liang, S. Poria, E. Cambria, L. Morency, 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2236–2246.

[10] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, D. Song, What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis, Inf. Fusion 66 (2021) 184–197.

[11] G. Chandrasekaran, T.N. Nguyen, J.H. D., Multimodal sentimental analysis for social media applications: A comprehensive review, WIREs Data Mining Knowl. Discov. 11 (5) (2021).

[12] S.A. Abdu, A.H. Yousef, A. Salem, Multimodal video sentiment analysis using deep learning approaches, a survey, Inf. Fusion 76 (2021) 204–226.

[13] V. Pérez-Rosas, R. Mihalcea, L. Morency, Utterance-level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers, The Association for Computer Linguistics, 2013, pp. 973–982.

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

[15] J.G. Ellis, B. Jou, S. Chang, Why we watch the news: A dataset for exploring sentiment in broadcast video news, in: A.A. Salah, J.F. Cohn, B.W. Schuller, O. Aran, L. Morency, P.R. Cohen (Eds.), Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, November 12-16, 2014, ACM, 2014, pp. 104–111.

[16] S. Park, H.S. Shim, M. Chatterjee, K. Sagae, L. Morency, Multimodal analysis and prediction of persuasiveness in online social multimedia, ACM Trans. Interact. Intell. Syst. 6 (3) (2016) 25:1–25:25.

[17] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, 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, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 3718–3727.

[18] A.B. Zadeh, Y. Cao, S. Hessner, P.P. Liang, S. Poria, L. Morency, CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 1801–1812.

[19] S. Poria, E. Cambria, A.F. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, The Association for Computational Linguistics, 2015, pp. 2539–2544.

[20] H. Wang, A. Meghawat, L. Morency, E.P. Xing, Select-additive learning: Improving generalization in multimodal sentiment analysis, in: 2017 IEEE International Conference on Multimedia and Expo, ICME 2017, Hong Kong, China, July 10-14, 2017, IEEE Computer Society, 2017, pp. 949–954.

[21] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, Association for Computational Linguistics, 2017, pp. 1103–1114.

[22] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2247–2256.

[23] S. Mai, H. Hu, S. Xing, Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 481–492.

[24] S. Mai, S. Xing, H. Hu, Locally confined modality fusion network with a global perspective for multimodal human affective computing, IEEE Trans. Multimed. 22 (1) (2020) 122–137.

[25] M. Chen, S. Wang, P.P. Liang, T. Baltrusaitis, A. Zadeh, L. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, United Kingdom, November 13 - 17, 2017, ACM, 2017, pp. 163–171.

[26] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L. Morency, Multi-attention recurrent network for human communication comprehension, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 5642–5649.

[27] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 5634–5641.

[28] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 7216–7223.

[29] Y. Wu, Y. Zhao, H. Yang, S. Chen, B. Qin, X. Cao, W. Zhao, Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors, in: Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 1397–1406.

[30] H. Pham, P.P. Liang, T. Manzini, L. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 6892–6899.

[31] Y.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 6558–6569.

[32] W. Rahman, M.K. Hasan, S. Lee, A.B. Zadeh, C. Mao, L. Morency, M.E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 2359–2369.

[33] Q. Zhang, L. Shi, P. Liu, Z. Zhu, L. Xu, ICDN: Integrating consistency and difference networks by transformer for multimodal sentiment analysis, Appl. Intell. (2022) 1–14.

[34] Z. Li, Y. Zhou, W. Zhang, Y. Liu, C. Yang, Z. Lian, S. Hu, AMOA: Global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis, in: Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, International Committee on Computational Linguistics, 2022, pp. 7136–7146.

[35] Z. Sun, P.K. Sarma, W.A. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 8992–8999.

[36] D. Hazarika, R. Zimmermann, S. Poria, MISA: Modality-invariant and -specific representations for multimodal sentiment analysis, in: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, ACM, 2020, pp. 1122–1131.

[37] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022).

[38] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, R. Xu, Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis, in: Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, International Committee on Computational Linguistics, 2022, pp. 7124–7135.

[39] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics, 2017, pp. 873–883.

[40] D. Ghosal, M.S. Akhtar, D.S. Chauhan, S. Poria, A. Ekbal, P. Bhattacharyya, Contextual inter-modal attention for multi-modal sentiment analysis, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 3454–3466.

[41] Y. Bengio, R. Ducharme, P. Vincent, A neural probabilistic language model, in: Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA, MIT Press, 2000, pp. 932–938.

[42] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008, in: ACM International Conference Proceeding Series, vol. 307, ACM, 2008, pp. 160–167.

[43] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013.

[44] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543.

[45] C.F. Benitez-Quiroz, Y. Wang, A.M. Martínez, Recognition of action units in the wild with deep nets and a new global-local loss, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 3990–3999.

[46] D. Tran, L.D. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015, pp. 4489–4497.

[47] G. Littlewort, J. Whitehill, T. Wu, I.R. Fasel, M.G. Frank, J.R. Movellan, M.S. Bartlett, The computer expression recognition toolbox (CERT), in: Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011, IEEE Computer Society, 2011, pp. 298–305.

[48] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L. Morency, OpenFace 2.0: Facial behavior analysis toolkit, in: 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018, IEEE Computer Society, 2018, pp. 59–66.

[49] A. Graves, S. Fernández, J. Schmidhuber, Bidirectional LSTM networks for improved phoneme classification and recognition, in: Artificial Neural Networks: Formal Models and their Applications - ICANN 2005, 15th International Conference, Warsaw, Poland, September 11-15, 2005, Proceedings, Part II, in: Lecture Notes in Computer Science, vol. 3697, Springer, 2005, pp. 799–804.

[50] F. Eyben, M. Wöllmer, A. Graves, B.W. Schuller, E. Douglas-Cowie, R. Cowie, On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues, J. Multimodal User Interfaces 3 (1–2) (2010) 7–19.

[51] N. Anand, P. Verma, Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data, 2015.

[52] F. Eyben, M. Wöllmer, B.W. Schuller, OpenEAR - Introducing the munich open-source emotion and affect recognition toolkit, in: Affective Computing and Intelligent Interaction, Third International Conference and Workshops, ACII 2009, Amsterdam, the Netherlands, September 10-12, 2009, Proceedings, IEEE Computer Society, 2009, pp. 1–6.

[53] F. Eyben, M. Wöllmer, B.W. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010, ACM, 2010, pp. 1459–1462.

[54] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, Librosa: Audio and music signal analysis in python, in: Proceedings of the 14th Python in Science Conference, Vol. 8, Citeseer, 2015, pp. 18–25.

[55] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP - A collaborative voice analysis repository for speech technologies, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, IEEE, 2014, pp. 960–964.

[56] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[57] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain, IEEE Computer Society, 2016, pp. 439–448.

[58] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 10790–10797.

[59] E. Shutova, D. Kiela, J. Maillard, Black holes and white rabbits: Metaphor identification with visual features, in: NAACL HLT 2016, the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, The Association for Computational Linguistics, 2016, pp. 160–170.

[60] E. Morvant, A. Habrard, S. Ayache, Majority vote of diverse classifiers for late fusion, in: Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings, in: Lecture Notes in Computer Science, vol. 8621, Springer, 2014, pp. 153–162.

[61] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention, IEEE Trans. Multimed. 15 (7) (2013) 1553–1568.

[62] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrusaitis, L. Morency, Deep multimodal fusion for persuasiveness prediction, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016, ACM, 2016, pp. 284–288.

[63] J. Kossaifi, Z.C. Lipton, A. Kolbeinsson, A. Khanna, T. Furlanello, A. Anandkumar, Tensor regression networks, J. Mach. Learn. Res. 21 (2020) 123:1–123:21.

[64] E.J. Barezi, P. Fung, Modality-based factorization for multimodal fusion, in: Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019, Association for Computational Linguistics, 2019, pp. 260–269.

[65] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, C.L. Giles, Learning to extract semantic structure from documents using multimodal fully convolutional neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 4342–4351.

[66] P.P. Liang, Z. Liu, Y.H. Tsai, Q. Zhao, R. Salakhutdinov, L. Morency, Learning representations from imperfect time series data via tensor rank regularization, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 1569–1576.

[67] P.P. Liang, Z. Liu, A. Zadeh, L. Morency, Multimodal language analysis with recurrent multistage fusion, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 150–161.

[68] Y. Gu, X. Li, K. Huang, S. Fu, K. Yang, S. Chen, M. Zhou, I. Marsic, Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder, in: 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018, ACM, 2018, pp. 537–545.

[69] Y. Wu, Z. Lin, Y. Zhao, B. Qin, L. Zhu, A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, in: Findings of ACL, vol. ACL/IJCNLP 2021, Association for Computational Linguistics, 2021, pp. 4730–4738.

[70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[71] F. Zhang, X. Li, C.P. Lim, Q. Hua, C. Dong, J. Zhai, Deep emotional arousal network for multimodal sentiment analysis and emotion recognition, Inf. Fusion 88 (2022) 296–304.

[72] D.R. Hardoon, S. Szedmák, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.

[73] Y. Zeng, Z. Li, Z. Tang, Z. Chen, H. Ma, Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis, Expert Syst. Appl. 213 (Part) (2023) 119240.

[74] N. Majumder, D. Hazarika, A.F. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl.-Based Syst. 161 (2018) 124–133.

[75] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, in: Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014, Association for Computational Linguistics, 2014, pp. 103–111.

[76] D.S. Chauhan, M.S. Akhtar, A. Ekbal, P. Bhattacharyya, Context-aware interactive attention for multi-modal sentiment and emotion analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 5646–5656.

[77] Z.C. Lipton, The mythos of model interpretability, ACM Queue 16 (3) (2018) 30.

[78] A. Holzinger, P. Kieseberg, E.R. Weippl, A.M. Tjoa, Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI, in: Machine Learning and Knowledge Extraction - Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27-30, 2018, Proceedings, in: Lecture Notes in Computer Science, vol. 11015, Springer, 2018, pp. 1–8.

[79] Y. Zhang, D. Song, P. Zhang, P. Wang, J. Li, X. Li, B. Wang, A quantum-inspired multimodal sentiment analysis framework, Theoret. Comput. Sci. 752 (2018) 21–40.

[80] Y. Zhang, D. Song, X. Li, P. Zhang, P. Wang, L. Rong, G. Yu, B. Wang, A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis, Inf. Fusion 62 (2020) 14–31.

[81] Q. Li, D. Gkoumas, C. Lioma, M. Melucci, Quantum-inspired multimodal fusion for video sentiment analysis, Inf. Fusion 65 (2021) 58–71.

[82] D. Borth, R. Ji, T. Chen, T.M. Breuel, S. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013, ACM, 2013, pp. 223–232.

[83] A. Khosla, A.D. Sarma, R. Hamid, What makes an image popular? in: 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, ACM, 2014, pp. 867–876.

[84] C. Schulze, D. Henter, D. Borth, A. Dengel, Automatic detection of CSA media by multi-modal feature fusion for law enforcement support, in: International Conference on Multimedia Retrieval, ICMR '14, Glasgow, United Kingdom - April 01 - 04, 2014, ACM, 2014, p. 353.

[85] M. Wankhade, A.C.S. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, Artif. Intell. Rev. 55 (7) (2022) 5731–5780.

[86] S. Madhu, An approach to analyze suicidal tendency in blogs and tweets using sentiment analysis, Int. J. Sci. Res. Comput. Sci. Eng. 6 (4) (2018) 34–36.

[87] T.K. Mackey, A. Miner, R.E. Cuomo, Exploring the e-cigarette e-commerce marketplace: Identifying Internet e-cigarette marketing characteristics and regulatory gaps, Drug Alcohol Depend. 156 (2015) 97–103.

[88] L. Zhu, M. Xu, Y. Xu, Z. Zhu, Y. Zhao, X. Kong, A multi-attribute decision making approach based on information extraction for real estate buyer profiling, World Wide Web (2022) 1–19.

[89] K.R. Apala, M. Jose, S. Motnam, C. Chan, K.J. Liszka, F. de Gregorio, Prediction of movies box office performance using social media, in: Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013, ACM, 2013, pp. 1209–1214.

[90] F.Z. Xing, E. Cambria, R.E. Welsch, Natural language based financial forecasting: a survey, Artif. Intell. Rev. 50 (1) (2018) 49–73.

[91] Y. Ma, R. Mao, Q. Lin, P. Wu, E. Cambria, Multi-source aggregated classification for stock price movement prediction, Inf. Fusion 91 (2023) 515–528.

[92] C. Langlet, C. Clavel, Grounding the detection of the user's likes and dislikes on the topic structure of human-agent interactions, Knowl.-Based Syst. 106 (2016) 116–124.

[93] J.G. Ellis, B. Jou, S. Chang, Why we watch the news: A dataset for exploring sentiment in broadcast video news, in: Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, November 12-16, 2014, ACM, 2014, pp. 104–111.

[94] B. Siddiquie, D. Chisholm, A. Divakaran, Exploiting multimodal affect and semantics to identify politically persuasive web videos, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015, ACM, 2015, pp. 203–210.

[95] C.N. Dang, M.N.M. García, F. de la Prieta, An approach to integrating sentiment analysis into recommender systems, Sensors 21 (16) (2021) 5666.

[96] R. Mao, Q. Liu, K. He, W. Li, E. Cambria, The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection, IEEE Trans. Affect. Comput. (2022).

[97] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, S. Poria, Towards multimodal sarcasm detection (An _obviously_ perfect paper), in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 4619–4629.

[98] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, in: Mining Text Data, Springer, 2012, pp. 415–463.

[99] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J. 5 (4) (2014) 1093–1113.

[100] S. Poria, A. Hussain, E. Cambria, Combining textual clues with audio-visual information for multimodal sentiment analysis, in: Multimodal Sentiment Analysis, Springer, 2018, pp. 153–178.

[101] R. Mao, X. Li, M. Ge, E. Cambria, MetaPro: A computational metaphor processing model for text pre-processing, Inf. Fusion 86–87 (2022) 30–43.

[102] J.S. Grosman, P.H.T. Furtado, A.M.B. Rodrigues, G.G. Schardong, S.D.J. Barbosa, H.C.V. Lopes, Eras: Improving the quality control in the annotation process for Natural Language Processing tasks, Inf. Syst. 93 (2020) 101553.

[103] R.H. Thiele, T.L. McMurry, Data agnosticism and implications on method comparison studies, Anesth. Analg. 121 (2) (2015) 264–266.

[104] D. Morrison, Toward automatic censorship detection in microblogs, in: Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2014 International Workshops: DANTH, BDM, MobiSocial, BigEC, CloudSD, MSMV-MBI, SDA, DMDA-Health, ALSIP, SocNet, DMBIH, BigPMA,Tainan, Taiwan, May 13-16, 2014. Revised Selected Papers, in: Lecture Notes in Computer Science, vol. 8643, Springer, 2014, pp. 572–583.

[105] B. Schuller, J.-G. Ganascia, L. Devillers, Multimodal sentiment analysis in the wild: Ethical considerations on data collection, annotation, and exploitation, in: Proceedings of the 1st International Workshop on ETHics in Corpus Collection, Annotation and Application (ETHI-CA 2016), Satellite of the 10th Language Resources and Evaluation Conference (LREC 2016)(2016), 2016, pp. 29–34.

[106] D. Zhang, S. Li, Q. Zhu, G. Zhou, Effective sentiment-relevant word selection for multi-modal sentiment analysis in spoken language, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 148–156.

[107] S. Han, R. Mao, E. Cambria, Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings, in: Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, International Committee on Computational Linguistics, 2022, pp. 94–104.

[108] M. Arjmand, M.J. Dousti, H. Moradi, TEASEL: A transformer-based speech-prefixed language model, 2021, CoRR abs/2109.05522.