

Multimodal Sentiment Analysis: A Survey

SONGNING LAI, the School of Information Science and Engineering in Shandong University, China

XIFENG HU, the School of Information Science and Engineering in Shandong University, China

HAOXUAN XU, the School of Information Science and Engineering in Shandong University, China

ZHAOXIA REN*, Assets and Laboratory Management Department in Shandong University, China

ZHI LIU†, the School of Information Science and Engineering in Shandong University, China

Multimodal sentiment analysis has become an important research area in the field of artificial intelligence. With the latest advances in deep learning, this technology has reached new heights. It has great potential for both application and research, making it a popular research topic. This review provides an overview of the definition, background, and development of multimodal sentiment analysis. It also covers recent datasets and advanced models, emphasizing the challenges and future prospects of this technology. Finally, it looks ahead to future research directions. It should be noted that this review provides constructive suggestions for promising research directions and building better performing multimodal sentiment analysis models, which can help researchers in this field.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Computer vision**; **Theory of mind**.

Additional Key Words and Phrases: Multimodal Sentiment Analysis, Multimodal Fusion, Affective Computing, Computer Vision

ACM Reference Format:

Songning Lai, Xifeng Hu, Haoxuan Xu, Zhaoxia Ren, and Zhi Liu. 2023. Multimodal Sentiment Analysis: A Survey. *J. ACM* 37, 4, Article 111 (August 2023), 17 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Emotion is a subjective reaction of an organism to external stimuli [9, 42]. Humans possess a powerful capacity for sentiment analysis, and researchers are currently exploring ways to make this ability available to artificial agents [19].

Sentiment analysis involves analyzing sentiment polarity through available information [3, 23]. With the rapid development of fields such as artificial intelligence, computer vision, and natural language processing, it is becoming increasingly possible for artificial agents to implement sentiment analysis. Sentiment analysis is an interdisciplinary research area that includes computer science, psychology, social science, and other fields [32, 40, 72].

*Corresponding author

†Corresponding author

Authors' addresses: Songning Lai, sonly@mail.sdu.edu.cn, the School of Information Science and Engineering in Shandong University, Qingdao, Shandong, China, 266000; Xifeng Hu, 201942544@mail.sdu.edu.cn, the School of Information Science and Engineering in Shandong University, Qingdao, Shandong, China, 266000; Haoxuan Xu, 202020120237@mail.sdu.edu.cn, the School of Information Science and Engineering in Shandong University, Qingdao, Shandong, China, 266000; Zhaoxia Ren, renzx@sdu.edu.cn, Assets and Laboratory Management Department in Shandong University, Qingdao, Shandong, China, 266000; Zhi Liu, liuzhi@sdu.edu.cn, the School of Information Science and Engineering in Shandong University, Qingdao, Shandong, China, 266000.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

Manuscript submitted to ACM

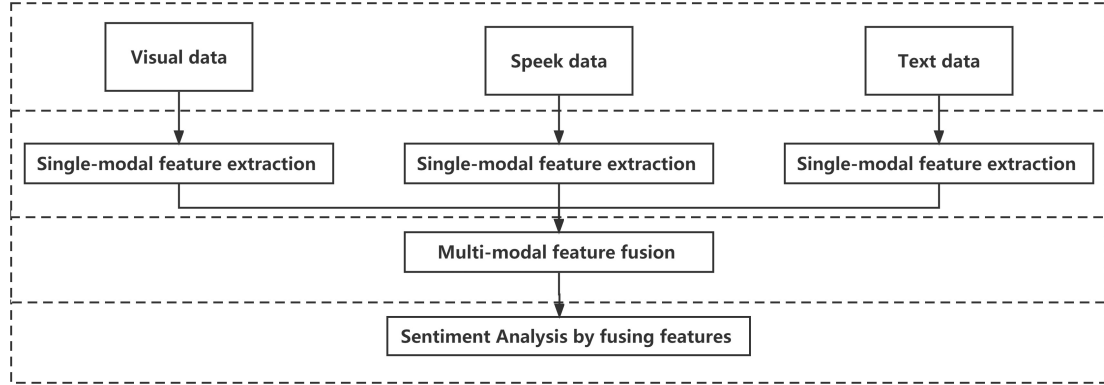


Fig. 1. Figure shows the model architecture for the more classical multi-modal sentiment analysis. The overall architecture consists of three parts: one part for feature extraction of individual modalities, one part for fusion of features of each modality, and one part for sentiment analysis of the fused features. These three parts are very important, and researchers have begun to optimize these three parts one by one

Scientists have been working to empower AI agents with sentiment analysis capabilities for decades. This is a key component of human-like AI, making AI more like a human.

Sentiment analysis has significant research value [4, 28, 59, 65]. With the explosive growth of Internet data, vendors can use evaluative data such as reviews and review videos to improve their products. Sentiment analysis also has countless research values, such as lie detection, interrogation, and entertainment. The following sections will elaborate on the application and research value of sentiment analysis.

In the past, sentiment analysis has mostly focused on a single modality (visual modality, speech modality, or text modality) [5]. Text-based sentiment analysis [25, 30, 53] has gone a long way in NLP. Vision-based sentiment analysis pays more attention to human facial expressions [8] and movement postures. Speech-based sentiment analysis mainly extracts features such as pitch, timbre, and temperament in speech for sentiment analysis [49]. With the development of deep learning, these three modalities have gained some foothold in sentiment analysis.

However, using a single modality for sentiment analysis has limitations [11, 17, 21, 26]. The emotional information contained in a single modality is limited and incomplete. Combining information from multiple modalities can provide deeper emotional polarity. Analyzing only one modality results in limited results and makes it difficult to accurately analyze the emotion of an action.

Researchers have gradually realized the need for multi-modal sentiment analysis, and many multi-modal sentiment analysis models have emerged to accomplish this task. Text features dominate and play a key role in the analysis of deep emotions[50]. Visual modality extraction of expression and pose features can effectively aid text sentiment analysis and judgment[29]. On the one hand, speech modality can extract text features, and on the other hand, speech tone can be recognized to reveal the state of text at each time point[10]. Figure1 shows the model architecture for the more classical multi-modal sentiment analysis. The overall architecture consists of three parts: one part for feature extraction of individual modalities, one part for fusion of features of each modality, and one part for sentiment analysis of the fused features. These three parts are very important, and researchers have begun to optimize these three parts one by one [43].

In this review, we provide a comprehensive overview of the field of multimodal sentiment analysis. The review includes a summary and brief introduction of datasets, which can help researchers select appropriate datasets. We compare and analyze models that have significant research significance in multimodal sentiment analysis and provide suggestions for model construction. We elaborate on three types of modal fusion methods and explain the advantages and disadvantages of different modal fusion methods. Finally, we look ahead to the challenges and future development directions of multimodal sentiment analysis, providing several promising research directions. Compared to other reviews in the same field, our focus is on providing constructive suggestions for promising research directions and building better performing multimodal sentiment analysis models. We emphasize the challenges and future prospects of these technologies.

2 MULTIMODAL SENTIMENT ANALYSIS DATASETS

With the growth of the Internet, an era of data explosion has been created[14, 37, 39]. Numerous researchers have widely collected these data from the Internet (videos, reviews, etc.) and built sentiment datasets according to their own needs. Tab1 summarizes the commonly used multimodal datasets. The first column indicates the name of the data set. The second column is the year in which the sentiment data was released. The third column is the category of modalities included in the sentiment dataset. The fourth column is the platform from which the data set came. The fifth column is the language used by the dataset. The sixth column is the amount of data contained in the dataset. Each dataset has its own characteristics. This section lists well-known datasets in the community, aiming to help researchers sort out the characteristics of each dataset and make it easier for them to choose datasets.

2.1 IEMOCAP[2]

IEMOCAP, a sentiment analysis dataset released by the Speech Analysis and Interpretation Laboratory in 2008, is a multi-modal dataset that comprises 1,039 conversational segments, with a total video length of 12 hours. Participants in the study engaged in five different scenarios, performing emotions as per a pre-set scenario. The dataset includes not only audio, video, and text information but also facial expression and posture information obtained through additional sensors. Data points are categorized into ten emotions: neutral, happy, sad, angry, surprised, scared, disgusted, frustrated, excited, and other. Overall, IEMOCAP provides a rich resource for researchers exploring sentiment analysis across multiple modalities.

2.2 DEAP[24]

DEAP is a dataset specifically designed for sentiment analysis that utilizes physiological signals as its source of data (Koelstra et al., 2011). The dataset examines EEG data from 32 subjects, with a 1:1 ratio of male and female participants. EEG signals were collected at 512Hz from different areas of the subjects' brains, including the frontal, parietal, occipital, and temporal lobes. To annotate the EEG signals, the subjects were asked to rate the corresponding videos in terms of three aspects: Valence, Arousal, and Dominance, on a scale of 1 to 9. This dataset provides valuable resources for researchers to explore sentiment analysis using physiological signals.

2.3 CMU-MOSI[70]

CMU-MOSI dataset is comprised of 93 critical YouTube videos that cover a range of topics (Zadeh et al., 2016). These videos were carefully selected to ensure that they featured only one speaker who was facing the camera, allowing for clear capture of facial expressions. While there were no restrictions on camera model, distance, or speaker scene,

Name	Year	Modalities	Source	Language	Number
IEMOCAP	2008	A+V+T	N/A	English	10039
DEAP	2011	A+V+T A+V+T A+V+T	N/A	English	10039
CMU-MOSI	2016	A+V+T	YouTube	English	2199
CMU-MOSEI	2018	A+V+T	YouTube	English	23453
MELD	2019	A+V+T	The Friends	English	13000
Multi-ZOL	2019	V+T	ZOL.com	Chinese	5288
CH-SIMS	2020	A+V+T	N/A	Chinese	2281
CMU-MOSEAS	2021	A+V+T	YouTube	Spanish Portuguese German French	40000
FACTIFY	2022	V+T	Twitter	English	50000
MEMOTION	2022	V+T	Reddit Facebook	English	10000

Table 1. This table contains the used multimodal datasets. The first column indicates the name of the data set. The second column is the year in which the sentiment data was released. The third column is the category of modalities included in the sentiment dataset. The fourth column is the platform from which the data set came. The fifth column is the language used by the dataset. The sixth column is the amount of data contained in the dataset.

all presentations and comments were made in English by 89 different speakers, including 41 women and 48 men. The 93 videos were divided into 2,199 subjective opinion segments and annotated with sentiment intensity ranging from strongly negative to strongly positive (-3 to 3). Overall, the CMU-MOSI dataset provides a valuable resource for researchers studying sentiment analysis.

2.4 CMU-MOSEI[71]

CMU-MOSEI is a popular dataset for sentiment analysis that comprises 3,228 YouTube videos (Zadeh et al., 2018). These videos are categorized into 23,453 segments and feature data from three different modalities: text, visual, and sound. With contributions from 1,000 speakers and coverage of 250 different topics, this dataset offers a diverse range of perspectives. All the videos are in English, and both sentiment and emotion annotations are available. The six emotion categories include happy, sad, angry, scared, disgusted, and surprised, while the sentiment intensity markers range from strongly negative to strongly positive (-3 to 3). Overall, CMU-MOSEI is an invaluable resource for researchers exploring sentiment analysis across multiple modalities.

2.5 MELD[45]

MELD is a comprehensive dataset that includes video clips from the popular television series Friends. The dataset comprises textual, audio, and video information that corresponds to the textual data. It contains 1400 videos, which are further divided into 13,000 individual segments. The dataset is annotated with seven categories of annotations: Anger,

Disgust, Sadness, Joy, Neutral, Surprise, and Fear. Each segment has three sentiment annotations: positive, negative, and neutral.

2.6 Multi-ZOL[63]

Multi-ZOL is a dataset designed for bimodal sentiment classification of images and text. The dataset consists of reviews of mobile phones collected from ZOL.com. It contains 5288 sets of multimodal data points that cover various models of mobile phones from multiple brands. These data points are annotated with a sentiment intensity rating from 1 to 10 for six aspects.

2.7 CH-SIMS[67]

CH-SIMS is a unique dataset consisting of 60 open-sourced videos from the web that are split into 2281 video clips. The dataset focuses solely on Chinese (Mandarin) language and ensures that each segment contains only one character's face and voice. It covers a wide range of scenes and speaker ages and is individually labeled for each modality, making it a valuable resource for researchers. The dataset annotations contain sentiment intensity ratings ranging from negative to positive (-1 to 1) and also include annotations for other attributes such as age and gender.

2.8 CMU-MOSEA[69]

CMU-MOSEA is a versatile dataset that includes multiple languages, such as Spanish, Portuguese, German, and French. The dataset comprises 40,000 sentence fragments, covering 250 different topics and 1645 speakers. The annotations are split into two categories: sentiment intensity and binary. Each sentence is annotated with sentiment strength in the interval [-3,3], and the binary includes whether the speaker expressed an opinion or made an objective statement. Emotions are divided into six categories for each sentence: happiness, sadness, fear, disgust, surprise, and annotation.

2.9 FACTIFY[35]

FACTIFY is a fake news detection dataset that focuses on implementation validation. It includes data for both image and text modalities and contains 50,000 sets of data. Most of the data's claims refer to politics and government. The dataset is annotated into three categories: support, no evidence, and refutation. This dataset is a valuable resource for researchers interested in detecting and combating the spread of fake news.

2.10 MEMOTION[47]

MEMOTION is a meme-based dataset that includes popular memes related to politics, religion, and sports. The dataset comprises 10,000 data points and is divided into three sub-tasks: Sentiment Analysis, Emotion Classification, and Scale/Intensity of Emotion Classes. The annotators annotated each data point differently under the different subtasks. Subtask one annotates each data point into three categories (negative, neutral, and positive). Subtask two annotates each data point into four categories (humour, sarcasm, offense, motivation). Subtask three annotates each data point in the interval [0,4] to indicate the sentiment intensity. This dataset provides a unique opportunity for researchers to analyze the use of memes as a means of communication and expression in modern culture.

The aforementioned datasets all have certain limitations. We summarize the limitations of each dataset to help researchers make informed decisions when selecting datasets for experiments. For IEMOCAP, the limited number of actors in the dataset may lead to overfitting issues. Additionally, the emotion categories in the dataset may not be comprehensive enough to cover all emotion types. For DEAP, CMU-MOSI, and CMU-MOSEI, the emotion categories

in the datasets may not be comprehensive enough to cover all emotion types. Furthermore, the video clips in the datasets may be too short to fully express emotions. For MELD, the audio samples in the dataset may be too short to fully express emotions. Additionally, facial expressions may be affected by environmental factors such as lighting. For Multi-ZOL, since the sample sources for social media comments are diverse, there may be some degree of noise. Furthermore, this dataset is only suitable for sentiment analysis tasks on social media comments. For CH-SIMS, this dataset is only suitable for student negative emotion recognition tasks. Additionally, due to the limited sample sources being restricted to student questionnaires, there may be a certain degree of subjective bias. For CMU-MOSEA, this dataset is only suitable for multimodal information and emotion recognition tasks in movie scenes. Additionally, due to the limited sample sources being restricted to movie scene annotations, there may be a certain degree of subjective bias. For FACTIFY, this dataset is only suitable for topic and sentiment recognition tasks in news articles. Due to the limited sample sources being restricted to news article annotations, there may be a certain degree of subjective bias.

3 MULTIMODAL FUSION

Multimodal data describe objects from different perspectives and are more informative than single-modal data. Data information from different modalities can be complementary to each other. In the task of multimodal sentiment analysis, it is a very important and challenging task to fuse the data features between different modalities, preserve the semantic integrity of the modalities, and achieve a good fusion between different modalities. According to the different modes of modal fusion, it can be summarized as feature-based multimodal fusion in the early stage, model-based multimodal fusion in the middle stage, and decision-based multimodal fusion in the late stage.

3.1 Early feature-based approaches for multimodal fusion

Early feature-based multimodal fusion methods perform shallow fusion after feature extraction in the early stage. The fusion of the features of different modalities at the shallow level of the model is equivalent to unifying the features of different single modalities into the same parameter space. Due to the differences in information between different modalities, features commonly contain a large amount of redundant information, and it is frequently necessary to use dimensionality reduction methods to remove the redundant information. Features after dimensionality reduction are fed into the model to complete feature extraction and prediction. Early feature fusion wants the model to consider input information from multiple modalities at the beginning of feature modeling. However, the method of unifying multiple different parameter spaces in the input layer may not achieve the desired effect due to the differences in the parameter spaces of different modes. This category of models can effectively handle multimodal emotion recognition tasks with high accuracy and robustness. However, these models require a large amount of training data to achieve good performance, and their structures are relatively complex, which requires a longer training time. The overall framework of early feature-based approaches for multimodal fusion is shown in Figure 2. Some representative models are:

3.1.1 THMM (Tri-modal Hidden Markov Model) [36]. One approach to multimodal sequence modeling and analysis is to represent eigenvectors of multiple modalities as higher-order tensors and use tensor decomposition methods to extract hidden states and transition probabilities. This approach effectively exploits the correlation and complementarity between multimodal data, while avoiding the curse of dimensionality and overfitting. However, the disadvantage is that the order and rank of the tensors and the number of hidden states must be predetermined, which may affect the model's performance and efficiency.

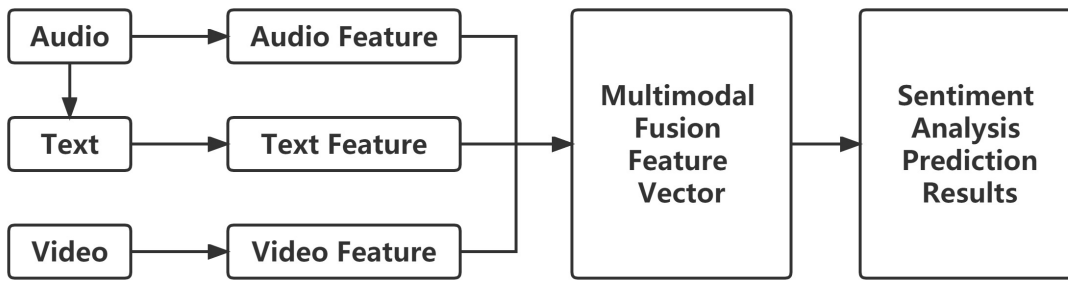


Fig. 2. Figure shows the overall framework of early feature-based approaches for multimodal fusion. After extracting features, this model uses specific components to fuse the features of each modality.

3.1.2 RMFN (Recurrent Multistage Fusion Network) [31]. Another approach is to use multiple recurrent neural network layers to gradually fuse features from different modalities, from local to global, from low-level to high-level, and finally obtain a comprehensive sentiment representation. This model uses an attention mechanism to adjust the location of features in semantic space for different modalities, allowing the same word to exhibit different emotions under different nonverbal behaviors.

3.1.3 RAVEN (Recurrent Attended Variation Embedding Network) [58]. A Hierarchical Fusion Network has been proposed for multimodal sentiment analysis, which includes a local fusion module and a global fusion module. Local cross-modal fusion is explored through a sliding window, which effectively reduces computational complexity.

3.1.4 HFFN (Hierarchical feature fusion network with local and global perspectives for multimodal affective computing) [34]. Another approach is to use recurrent neural networks and adversarial learning to learn joint representations between different modalities, thereby improving the ability of single-modal representations and dealing with missing modalities or noise.

3.1.5 MCTN (Multimodal Cyclic Translation Network) [41]. A medium-term model-based multimodal fusion approach involves feeding multimodal data into the network, and the intermediate layers of the model perform feature fusion between the modalities. Model-based modality fusion methods can select the location of modality feature fusion to achieve intermediate interactions. Model-based fusion typically uses multiple kernel learning, neural networks, graph models, and alternative methods.

3.2 Medium-term model-based multimodal fusion method

A medium-term model-based multimodal fusion approach is to feed multimodal data into the network, and the intermediate layers of the model perform feature fusion between the modalities. Model-based modality fusion methods can select the location of modality feature fusion to achieve intermediate interactions. Model-based fusion typically uses multiple kernel learning, neural networks, graph models, and alternative methods. The overall framework of medium-term model-based multimodal fusion method is shown in Figure 3.

3.2.1 MKL (Multiple Kernel Learning) [44]. This model is a multiple kernel learning approach. It uses different kernel functions to represent different modal information and selects the optimal combination of kernel functions by

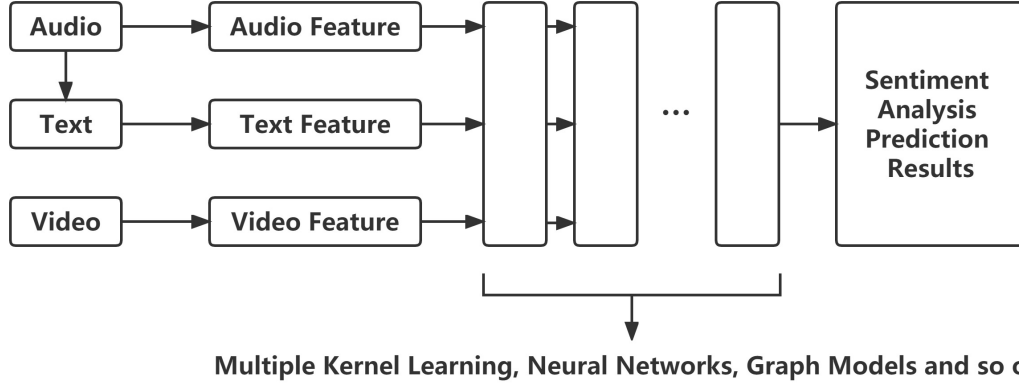


Fig. 3. Figure shows the overall framework of medium-term model-based approaches for multimodal fusion. This class inputs the feature information of each modality into multiple kernel learning, neural networks, graph models, and alternative methods to complete the fusion of modalities. Most of the nodes of its modal fusion are variable.

optimizing an objective function to achieve the fusion of multi-modal information. The model has high flexibility and can adaptively select different kernel functions, thereby improving the robustness and accuracy of the model. However, the model structure is relatively simple and may not be able to handle complex emotional expressions.

3.2.2 BERT-like (Self Supervised Models to Improve Multimodal Speech Emotion Recognition) [51]. This model is a Transformer-based multi-modal sentiment analysis method that can leverage self-attention mechanism to achieve alignment and fusion between text and image. The model adopts a self-supervised learning method, which can effectively handle multimodal emotion recognition tasks with high accuracy and robustness. In addition, the model may be affected by the quality of data and annotations.

3.3 Multimodal model based on decision fusion in the later stage

A decision level fusion method is used to fuse information from different modalities. Decision-level fusion refers to training models separately on data from different modalities to incorporate outputs from different modalities into the final decision. Multimodal models based on decision fusion typically fuse modalities using methods such as averaging, majority voting, weighting, and learnable models. Such models are typically lightweight and flexible. When any modality is missing, the decision can be made by using the remaining modalities. The overall framework of multimodal model based on decision fusion in the later stage method is shown in Figure 4.

3.3.1 Deep Multimodal Fusion Architecture [38]. In this model, each modality has an independent classifier. The prediction results are output after averaging the confidence scores of each classifier. The model has a simple structure, is easy to implement, and can effectively handle multimodal emotion recognition tasks. However, the model cannot handle the interaction information between modalities and may suffer from information loss.

3.3.2 SAL-CNN (Select-Additive Learning CNN) [57]. This model is a multimodal sentiment analysis model based on CNN and attention mechanism. It uses an adaptive attention mechanism to fuse text and image features, a spatial attention mechanism to extract text-related regions in images, and finally a completely connected layer to classify

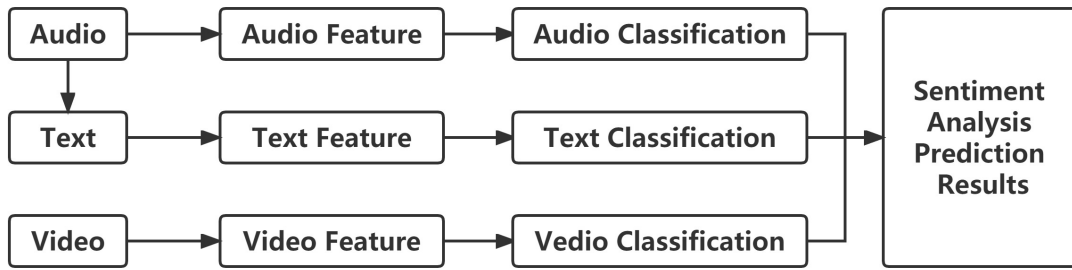


Fig. 4. Figure shows the overall framework of multimodal model based on decision fusion in the later stages method. These models train a separate classifier for each modality. Finally, the classifiers of each modality were integrated to complete the task of multimodal sentiment analysis.

the output. The model adopts an attention mechanism, which can effectively handle multimodal emotion recognition tasks with high accuracy and robustness. However, the model requires a large amount of training data to achieve good performance, and the model structure is relatively complex, requiring a longer training time.

3.3.3 TSAM (Temporally Selective Attention Model) [66]. The proposed model is a time-selective attention model, which assigns weights through an attention mechanism to help the model choose the time step, and finally sends it to a distinct SDL loss function model for sentiment analysis. SDL is a multi-modal sentiment analysis method based on Self-Distillation Learning, which can exploit the complementarity between different modalities to improve the generalization ability and robustness of the model. The model adopts a time-selective attention mechanism, which can effectively handle multimodal emotion recognition tasks with high accuracy and robustness.

4 LATEST MULTIMODAL SENTIMENT ANALYSIS MODELS

In recent years, the field of multimodal sentiment analysis has evolved into a huge system and many practical and efficient models and architectures have emerged. It's hard to cover all the models here. In this chapter, we present some recent and cutting-edge multimodal sentiment analysis models. Most of these models were used as benchmark models by later researchers for experimental reference. These models are summarized in Tab2. The first column is the name of the model. The second column is the year in which the model was published. The third column is the dataset used by the model. The fourth column is the accuracy under this dataset.

4.0.1 MultiSentiNet-Att[62]. This model uses an LSTM network to incorporate text information into word vectors. VGG is used to extract both target feature information and scene feature information of an image. The target and scene feature vectors are used to perform cross-modal attention mechanism learning with word vectors. That is, the target feature information and the scene feature information of the image are combined to assign special weights to the word vectors related to the sentiment in the text. The resulting features are fed into a multi-layer perceptron to complete the sentiment analysis task.

4.0.2 DFF-ATMF[7]. The proposed model mainly considers text modality and audio modality. The main contribution is to propose new multi-feature fusion strategies and multi-modal fusion strategies. Two parallel branches are used to

Name	Year	Dataset	Acc
MultiSentiNet-Att	2017	MVSA	68.86%
DFF-TMF	2019	CMU-MOSI	80.98%
		CMU-MOSEI	77.15%
AHRM	2020	Flickr	87.10%
		Getty Image	87.80%
SFNN	2020	Yelp	62.90%
MISA	2020	MOSI	83.40%
MAG-BERT	2020	CMU-MOSI	84.10%
		CMU-MOSEI	84.50%
TIMF	2021	CMU-MOSI	92.28%
		CMU-MOSEI	79.46%
Auto-ML based Fusion	2021	B-T4SA	95.19%
Self-MM	2022	CMU-MOSI	84.00%
		CMU-MOSEI	82.81%
		CH-SIMS	80.74%
DISRFN	2022	CMU-MOSI	83.60%
		CMU-MOSEI	87.50%

Table 2. This table contains some of the most recent and top-performing multimodal sentiment analysis models. The first column is the name of the model. The second column is the year in which the model was published. The third column is the dataset used by the model. The fourth column is the accuracy under this dataset.

learn features for text modality and features for audio modality. For the features of these two modalities, a multimodal attention fusion module is used to complete the multimodal fusion.

4.0.3 AHRM[61]. This model is mainly used to capture the relationship between text modality and visual modality. The authors propose an attention mechanism based heterogeneous relation model, which can well integrate the respective high-quality representation information of text modality and visual modality. This progressive dual attention mechanism can well highlight the channel-level semantic information of image and text information. To integrate social attributes, social relations are introduced to capture complementary information from the social context, and heterogeneous networks are constructed to integrate features.

4.0.4 SFNN[60]. The proposed model is a neural network based on semantic feature fusion. Convolutional neural networks and attention mechanisms are used to extract visual features. Visual features are mapped to text features and combined with text modality features for sentiment analysis.

4.0.5 MISA[15]. The proposed model presents a novel multi-modal sentiment analysis framework. Each modality is mapped into two distinct feature spaces after feature extraction. One feature space mainly learns the invariant features of the modality and the other one learns the unique features of the modality.

4.0.6 **MAG-BERT**[46]. The authors propose a "multi-modal" adaptation architecture and apply it to BERT. The model can receive input from multiple modalities during fine-tuning. MAG can be thought of as a vector embedding structure that allows us to input multimodal information and embed it as a sequence to BERT.

4.0.7 **TIMF**[54]. The main idea of this model is that each modality learns features separately and performs tensor fusion of the features of each modality. In the dataset fusion stage, the feature fusion for each modality is implemented by a tensor fusion network. In the decision fusion stage, the upstream results are fused by soft fusion to adjust the decision results.

4.0.8 **Auto-ML based Fusion**[33]. The authors propose to combine text and image individual sentiment analysis into a final fused classification based on AutoML. This approach combines individual classifiers into a final classification using the best model generated by Auto-ML. This is a typical model for decision-level fusion.

4.0.9 **Self-MM**[68]. In, the authors combine self-supervised learning and multi-task learning to construct a novel multi-modal sentiment analysis architecture. To learn the private information of each modality, the authors construct a single-modal label generation module ULGM based on self-supervised learning. The loss function corresponding to this module is designed to incorporate the private features learned by the three self-supervised learning subtasks into the original multi-modal sentiment analysis model using a weight adjustment strategy. The proposed model performs well, and the self-supervised learning based ULGM module also has the ability of single-modal label calibration.

4.0.10 **DISRFN**[16]. The model is a dynamically invariant representation-specific fusion network. The joint domain separation network is improved to obtain a joint domain separation representation for all modalities, so that redundant information can be effectively utilized. Second, a HGFN network is used to dynamically fuse the feature information of each modality and learn the features of multiple modal interactions. At the same time, a loss function that improves the fusion effect is constructed to help the model learn the representation information of each modality in the subspace.

5 MODEL COMPARISON AND SUGGESTIONS

This section evaluates five state-of-the-art multimodal sentiment analysis models: DFF-ATMF, MAG-BERT, TIMF, Self-MM, and DISRFN. While DFF-ATMF does not consider the vision modality, the other models analyze sentiment from three modalities of audio, text, and vision.

For the interaction relations of multimodal data, DFF-ATMF and TIMF build transformer-based models to learn complex relationships among the data. MAG-BERT uses a simple yet effective multimodal adaptive gate fusion strategy. Self-MM uses self-supervised multi-task learning as the fusion strategy, self-supervised generation of single-modal labels, and combination of single-modal labels to complete the multi-modal sentiment analysis task. DISRFN uses a Dynamic Invariant-Specific Representation Fusion Network to obtain jointly domain-separated representations of all modalities and dynamically fuse each representation through a hierarchical graph fusion network.

DFF-ATMF uses two parallel branches to fuse audio and text modalities. Its core mechanisms are feature vector fusion and multimodal attention fusion, which can learn more comprehensive sentiment information. However, due to the use of multi-layer neural networks and sophisticated fusion methods, overfitting may occur. Advantages: Simple structure, easy to implement, and can learn comprehensive sentiment information through feature vector fusion and multimodal attention fusion. Disadvantages: Does not consider the vision modality, may suffer from overfitting due to the use of multi-layer neural networks and sophisticated fusion methods.

MAG-BERT adapts the interior of BERTs using multimodal adaptation gates, which employ a simple yet effective fusion strategy without changing the structure and parameters of BERTs. However, the multimodal attention can only be performed within the same timestep but not across timesteps, which may ignore some temporal relationships. Additionally, MAG-BERT requires freezing the parameters of BERT without being able to fine-tune BERT, which may result in a representation of BERT that is not adapted to a specific task or domain. Advantages: Uses a simple yet effective multimodal adaptive gate fusion strategy without changing the structure and parameters of BERTs. Disadvantages: Multimodal attention can only be performed within the same timestep but not across timesteps, which may ignore some temporal relationships. Requires freezing the parameters of BERT without being able to fine-tune BERT, which may result in a representation of BERT that is not adapted to a specific task or domain.

TIMF leverages the self-attention mechanism of Transformers to learn complex interactions between multimodal data and generate unified sentiment representations. While it has the advantage of being able to learn complex relationships between modalities, it may suffer from extreme computational complexity, long training times, and problems with large amounts of labeled data. Advantages: Leverages the self-attention mechanism of Transformers to learn complex interactions between multimodal data and generate unified sentiment representations. Disadvantages: May suffer from extreme computational complexity, long training times, and problems with large amounts of labeled data.

Model	CMU-MOSI				CMU-MOSEI			
	MAE	Corr	Acc	F1-Score	MAE	Corr	Acc	F1-Score
DFF-ATMF	–	–	80.9	81.3	–	–	77.2	78.3
MAG-BERT	0.712	0.796	–	86	0.623	0.677	82	82.1
TIMF	0.373	0.93	92.3	92.3	0.645	0.669	79.5	79.5
Self-MM	0.723	0.797	84.8	84.8	0.534	0.764	84.1	84.1
DISRFN	0.798	0.734	83.4	83.6	0.591	0.78	87.5	87.5

Table 3. The table shows the performance metrics of the DFF-ATMF, MAG-BERT, TIMF, Self-MM and DISRFN models under the CMU-MOSI and CMU-MOSEI datasets. The evaluation parameters included: MAE, Corr, Acc and F1-Score.

Self-MM is a self-supervised multi-modal sentiment analysis model that uses a multi-task learning strategy to learn both multimodal and unimodal emotion recognition tasks. Its advantage is that it can generate single-modal labels using a self-supervised approach, saving the cost and time of manual labeling. However, interference and imbalance between multiple tasks can occur, and an appropriate weight adjustment strategy needs to be designed to balance the learning progress of different tasks. Advantages: Self-supervised multi-task learning as the fusion strategy, self-supervised generation of single-modal labels, and combination of single-modal labels to complete the multi-modal sentiment analysis task. Disadvantages: May require a large amount of labeled data to achieve good performance, and the self-supervised learning process may be computationally expensive.

DISRFN is a deep residual network-based multi-modal sentiment analysis model that exploits the strategy of Dynamic Invariant-Specific Representation Fusion Network to improve sentiment recognition capability. Its advantage is that it can efficiently utilize redundant information to obtain joint domain-separated representations of all modalities through a modified joint domain separation network and dynamically fuse each representation through a hierarchical graph fusion network to obtain the interaction information of multimodal data. However, as with Self-MM, interference and imbalance between multiple tasks can occur, and a suitable weight adjustment strategy needs to be designed to balance the learning progress of different tasks. Advantages: Uses a Dynamic Invariant-Specific Representation Fusion

Network to obtain jointly domain-separated representations of all modalities and dynamically fuse each representation through a hierarchical graph fusion network. Disadvantages: May require a large amount of labeled data to achieve good performance, and the hierarchical graph fusion network may be computationally expensive.

Tab3 shows the performance metrics of these five models under the CMU-MOSI and CMU-MOSEI datasets. When analyzing the performance metrics of the five models on these datasets, we recommend using BERT to extract features of text information while using LSTM to extract features for video and audio modality information since it requires capturing modality information in the time series.

DFF-ATMF does not consider visual modality, resulting in relatively low performance metrics. Visual information can provide additional information about human expressions, poses, scenes, etc., which can enhance the information of text and speech modalities as well as complement them. Therefore, visual modality information deserves to be considered and explored in multimodal sentiment analysis.

6 CHALLENGES AND FUTURE SCOPE

With the development of deep learning, multimodal sentiment analysis techniques have also been rapidly developed[18, 20, 22, 52]. However, multi-modal sentiment analysis still faces many challenges. This subsection analyzes the current state of research, challenges, and future developments in multimodal sentiment analysis.

6.1 Dataset

In multimodal sentiment analysis, the dataset plays a crucial role. Currently, a large dataset in multiple languages is missing. Given the diversity of languages and races in many countries, a large, diverse dataset could be used to train a multi-modal sentiment analysis model with strong generalization and wide usage. Additionally, current multimodal datasets still have low annotation accuracy and have not yet reached absolute continuous values, requiring researchers to label multimodal datasets more finely. Most current multimodal data contain only visual, speech, and text modalities and lack modal information combined with physiological signals such as brain waves and pulses.

6.2 Detection of Hidden Emotions

There has always been a recognized difficulty in multimodal sentiment analysis tasks: the analysis of hidden emotions. Hidden emotions[13, 55] include: sarcastic emotions (such as sarcastic words), emotions that need to be concretely analyzed in context, and complex emotions[27, 48] (such as a person's happiness and sadness). It is important to explore these hidden emotions. It's the gap between human and artificial intelligence[12].

6.3 Multiple forms of video data

In multimodal sentiment analysis tasks, video data is particularly challenging. Although the speaker is facing the camera and the video data resolution is maintained at a high level, the actual situation is more complicated and requires the model to be robust against noise and applicable to low-resolution video data. Capturing the micro-expressions and micro-gestures of speakers for sentiment analysis is also an area worth exploring by researchers.

6.4 Multiform language data

The form of text data in multimodal sentiment analysis tasks is typically single. However, evaluation texts in online communities are often cross-lingual, with reviewers using multiple languages to make more vivid comments. Text data with mixed emotions also remains a challenge for multimodal sentiment analysis tasks. Making good use of memes

that are mixed in the text is an important research topic, as memes often contain extremely strong emotional messages about reviewers. Additionally, most text data is transcribed directly through speech, making it particularly difficult to analyze a person's emotions when multiple people are talking. Combined with the cultural characteristics of different regions and countries, the same text data may reflect different emotions.

6.5 Future Prospects

The future of multimodal sentiment analysis techniques is extremely bright, and some of the future applications are listed below. Multimodal emotion analysis for real-time assessment of mental health[1, 56, 64]; multimodal criminal linguistic deception detection model[6]; offensive language detection; A human-like emotion-aware robot, etc. Multimodal emotion analysis is a technique for recognizing and analyzing emotions. Models that combine multi-modal information data for sentiment analysis can effectively improve the accuracy of sentiment analysis. In the future, multi-modal sentiment analysis techniques will be gradually improved. Perhaps one day there will be a multimodal sentiment analysis model with a large number of parameters that will have the same sentiment analysis capabilities as humans. It was a thing of rapture.

7 CONCLUSION

Multimodal sentiment analysis techniques have been recognized as important by researchers in various fields, making it a central research topic in the fields of natural language processing and computer vision. In this review, we provide a detailed description of various aspects of multimodal sentiment analysis, including its research background, definition, and development process. We also summarize commonly used benchmark datasets in Table 1 and compare and analyze recent state-of-the-art multimodal sentiment analysis models. Finally, we present the challenges posed by the field of multimodal sentiment analysis and explore possible future developments.

Many prospective works are being actively carried out and have even been largely implemented. However, there are still challenges to be addressed, leading to the following meaningful research directions:

- (1) Construct a large multimodal sentiment dataset in multiple languages.
- (2) Solve the domain transfer problem of video, text, and speech modal data.
- (3) Build a unified, large-scale multimodal sentiment analysis model with excellent generalization performance.
- (4) Reduce model parameters, optimize algorithms, and reduce algorithm complexity.
- (5) Solve the multi-lingual hybridness problem in multimodal sentiment analysis.
- (6) Discuss the weight problem of modal fusion and provide the most reasonable scheme to assign weights of different modalities in different cases.
- (7) Discuss the correlation between modalities and separate shared and private information between them to improve model performance and interpretability.
- (8) Construct a multimodal sentiment analysis model that can well complete hidden emotions.

DECLARATIONS

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by Joint found for smart computing of Shandong Natural Science Foundation under Grant ZR2020LZH013; open project of State Key Laboratory of Computer Architecture CARCHA202001; the Major Scientific and Technological Innovation Project in Shandong Province under Grant 2021CXG010506 and 2022CXG010504; "New University 20 items" Funding Project of Jinan under Grant 2021GXRC108 and 2021GXRC024.

Acknowledgments

Not applicable.

REFERENCES

- [1] Nujud Alosbhan, Anna Esposito, and Alessandro Vinciarelli. 2022. What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech. *Cognitive Computation* 14, 5 (2022), 1585–1598.
- [2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42 (2008), 335–359.
- [3] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems* 28, 2 (2013), 15–21.
- [4] Jireh Yi-Le Chan, Khean Thye Bea, Steven Mun Hong Leow, Seuk Wai Phoong, and Wai Khuen Cheng. 2023. State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review* 56, 1 (2023), 749–780.
- [5] Ganesh Chandrasekaran, Tu N Nguyen, and Jude Hemanth D. 2021. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, 5 (2021), e1415.
- [6] Safa Chebbi and Sofia Ben Jebara. 2021. Deception detection using multimodal fusion approaches. *Multimedia Tools and Applications* (2021), 1–30.
- [7] Feiyang Chen, Ziqian Luo, Yanyan Xu, and Dengfeng Ke. 2019. Complementary fusion of multi-features and multi-modalities in sentiment analysis. *arXiv preprint arXiv:1904.08138* (2019).
- [8] Rong Dai. 2020. Facial expression recognition method based on facial physiological features and deep learning. *Journal of Chongqing University of Technology (Natural Science)* 34, 6 (2020), 146–153.
- [9] Julien Deonna and Fabrice Teroni. 2012. *The emotions: A philosophical introduction*. Routledge.
- [10] Anna Favaro, Chelsie Motley, Tianyu Cao, Miguel Iglesias, Ankur Butala, Esther S Oh, Robert D Stevens, Jesús Villalba, Najim Dehak, and Laureano Moro-Velázquez. 2023. A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 532–539.
- [11] Ankita Gandhi, Kinjal Adharyu, and Vidhi Khanduja. 2021. Multimodal sentiment analysis: review, application domains and future directions. In *2021 IEEE Pune Section International Conference (PuneCon)*. IEEE, 1–5.
- [12] Fereshteh Ghanbari-Adivi and Mohammad Mosleh. 2019. Text emotion detection in social networks using a novel ensemble classifier based on Parzen Tree Estimator (TPE). *Neural Computing and Applications* 31, 12 (2019), 8971–8983.
- [13] Petr Hajek, Aliaksandr Barushka, and Michal Munk. 2020. Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Computing and Applications* 32 (2020), 17259–17274.
- [14] Abdul Haseeb, Enjun Xia, Shah Saud, Ashfaq Ahmad, and Hamid Khurshid. 2019. Does information and communication technologies improve environmental quality in the era of globalization? An empirical analysis. *Environmental Science and Pollution Research* 26 (2019), 8594–8608.
- [15] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1122–1131.
- [16] Jing He, Haonan Yanga, Changfan Zhang, Hongrun Chen, and Yifu Xua. 2022. Dynamic invariant-specific representation fusion network for multimodal sentiment analysis. *Computational Intelligence and Neuroscience* 2022 (2022).
- [17] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems* 167 (2019), 26–37.
- [18] Mahesh G Huddar, Sanjeev S Sannakki, and Vijay S Rajpurohit. 2019. A survey of computational approaches and challenges in multimodal sentiment analysis. *Int. J. Comput. Sci. Eng* 7, 1 (2019), 876–883.
- [19] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.
- [20] Anurag Illendula and Amit Sheth. 2019. Multimodal emotion classification. In *companion proceedings of the 2019 world wide web conference*. 439–449.

- [21] LIU Jiming, ZHANG Peixiang, LIU Ying, ZHANG Weidong, and FANG Jie. 2021. Summary of multi-modal sentiment analysis technology. *Journal of Frontiers of Computer Science & Technology* 15, 7 (2021), 1165.
- [22] Ramandeep Kaur and Sandeep Kautish. 2022. Multimodal sentiment analysis: A survey and comparison. *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines* (2022), 1846–1870.
- [23] Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. 1367–1373.
- [24] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.
- [25] Bernhard Kratzwald, Suzana Ilic, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Decision support with text-based emotion recognition: Deep learning for affective computing. *arXiv preprint arXiv:1803.06397* (2018).
- [26] Akshi Kumar and Geetanjali Garg. 2019. Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications* 78 (2019), 24103–24119.
- [27] Soonil Kwon. 2019. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* 20, 1 (2019), 183.
- [28] Hui Li, Qi Chen, Zhaoman Zhong, Rongrong Gong, and Guokai Han. 2022. E-word of mouth sentiment analysis for user behavior studies. *Information Processing & Management* 59, 1 (2022), 102784.
- [29] Jia Li, Ziyang Zhang, Junjie Lang, Yueqi Jiang, Liuwei An, Peng Zou, Yangyang Xu, Sheng Gao, Jie Lin, Chunxiao Fan, et al. 2022. Hybrid Multimodal Feature Extraction, Mining and Fusion for Sentiment Analysis. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 81–88.
- [30] Yang Li, Quan Pan, Suhang Wang, Tao Yang, and Erik Cambria. 2018. A generative model for category text generation. *Information Sciences* 450 (2018), 301–315.
- [31] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920* (2018).
- [32] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.
- [33] Vasco Lopes, António Gaspar, Luis A Alexandre, and João Cordeiro. 2021. An AutoML-based approach to multimodal image sentiment analysis. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–9.
- [34] Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 481–492.
- [35] Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Factify: A multi-modal fact verification dataset. In *Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY)*.
- [36] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. 169–176.
- [37] Sathyan Munirathinam. 2020. Industry 4.0: Industrial internet of things (IIOT). In *Advances in computers*. Vol. 117. Elsevier, 129–164.
- [38] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 284–288.
- [39] Esteban Ortiz-Ospina and Max Roser. 2023. The rise of social media. *Our world in data* (2023).
- [40] Arshi Parvaiz, Muhammad Anwaar Khalid, Rukhsana Zafar, Huma Ameer, Muhammad Ali, and Muhammad Moazam Fraz. 2023. Vision transformers in medical computer vision—A contemplative retrospection. *Engineering Applications of Artificial Intelligence* 122 (2023), 106126.
- [41] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6892–6899.
- [42] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.
- [43] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion* 37 (2017), 98–125.
- [44] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. 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*. 2539–2544.
- [45] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).
- [46] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2020. NIH Public Access, 2359.
- [47] Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR.
- [48] Umar Rashid, Muhammad Waseem Iqbal, Muhammad Akmal Skiandar, Muhammad Qasim Raiz, Muhammad Raza Naqvi, and Syed Khuram Shahzad. 2020. Emotion detection of contextual text using deep learning. In *2020 4th International symposium on multidisciplinary studies and innovative*

- technologies (ISMSIT). IEEE, 1–5.
- [49] Zhu Ren, Jia Jia, Quan Guo, Kuo Zhang, and Lianhong Cai. 2014. Acoustics, content and geo-information based sentiment prediction from large-scale networked voice data. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–4.
- [50] Vaibhav Rupapara, Furqan Rustam, Hina Fatima Shahzad, Arif Mehmood, Imran Ashraf, and Gyu Sang Choi. 2021. Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access* 9 (2021), 78621–78634.
- [51] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv preprint arXiv:2008.06682* (2020).
- [52] Lukas Stappen, Alice Baird, Lea Schumann, and Schuller Bjorn. 2021. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing* (2021).
- [53] Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*. 70–74.
- [54] Jianguo Sun, Hanqi Yin, Ye Tian, Junpeng Wu, Linshan Shen, and Lei Chen. 2021. Two-level multimodal fusion for sentiment analysis in public security. *Security and Communication Networks* 2021 (2021), 1–10.
- [55] Donglei Tang, Zhikai Zhang, Yulan He, Chao Lin, and Deyu Zhou. 2019. Hidden topic-emotion transition model for multi-level social emotion detection. *Knowledge-Based Systems* 164 (2019), 426–435.
- [56] Rahee Walambe, Pranav Nayak, Ashmit Bhardwaj, and Ketan Kotecha. 2021. Employing multimodal machine learning for stress detection. *Journal of Healthcare Engineering* 2021 (2021), 1–12.
- [57] Haoan Wang, Aaksha Meghawati, Louis-Philippe Morency, and Eric P Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 949–954.
- [58] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7216–7223.
- [59] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* 55, 7 (2022), 5731–5780.
- [60] Weidong Wu, Yabo Wang, Shuning Xu, and Kaibo Yan. 2020. SFNN: Semantic Features Fusion Neural Network for Multimodal Sentiment Analysis. In *2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE)*. IEEE, 661–665.
- [61] Jie Xu, Zhoujun Li, Feiran Huang, Chaozhuo Li, and S Yu Philip. 2020. Social image sentiment analysis by exploiting multimodal content and heterogeneous relations. *IEEE Transactions on Industrial Informatics* 17, 4 (2020), 2974–2982.
- [62] Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2399–2402.
- [63] Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 371–378.
- [64] Zhentao Xu, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Inferring social media users' mental health status from multimodal information. In *Proceedings of the 12th language resources and evaluation conference*. 6292–6299.
- [65] Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review* 53, 6 (2020), 4335–4385.
- [66] Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. 2017. Temporally selective attention model for social and affective state recognition in multimedia content. In *Proceedings of the 25th ACM international conference on Multimedia*. 1743–1751.
- [67] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3718–3727.
- [68] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10790–10797.
- [69] Amir Zadeh, Yan Sheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2020. CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Conference on Empirical Methods in Natural Language Processing, Vol. 2020. NIH Public Access, 1801.
- [70] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).
- [71] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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 (Volume 1: Long Papers)*. 2236–2246.
- [72] Bo Zhang, Jun Zhu, and Hang Su. 2023. Toward the third generation artificial intelligence. *Science China Information Sciences* 66, 2 (2023), 1–19.

Received 20 June 2023; revised 12 March 2009; accepted 5 June 2009