SCIENCE CHINA
Information Sciences

• REVIEW •

October 2025, Vol. 68, Iss. 10, 200101:1–200101:29
https://doi.org/10.1007/s11432-024-4593-8

Special Topic: Large Multimodal Models

Large language models meet text-centric multimodal sentiment analysis: a survey

Hao YANG[1], Yanyan ZHAO[1*], Yang WU[1], Shilong WANG[1], Tian ZHENG[1], Hongbo ZHANG[1], Zongyang MA[2], Wanxiang CHE[1], Shijin WANG[3], Si WEI[3*] & Bing QIN[1]

[1]*Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China*
[2]*Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[3]*iFLYTEK Co., Ltd., Hefei 230088, China*

Received 29 June 2024/Revised 5 March 2025/Accepted 8 May 2025/Published online 28 September 2025

Abstract Compared to traditional sentiment analysis, which only considers text, multimodal sentiment analysis needs to consider emotional signals from multimodal sources simultaneously and is therefore more consistent with the way how humans process sentiment in real-world scenarios. It involves processing emotional information from various sources such as natural language, images, videos, audio, and physiological signals. However, although other modalities also contain diverse emotional cues, natural language usually contains richer contextual information and therefore always occupies a crucial position in multimodal sentiment analysis. The emergence of ChatGPT has opened up immense potential for applying large language models (LLMs) to text-centric multimodal tasks. However, it is still unclear how existing LLMs can adapt better to text-centric multimodal sentiment analysis tasks. This survey aims to (1) present a comprehensive review of recent research in text-centric multimodal sentiment analysis tasks, (2) examine the potential of LLMs for text-centric multimodal sentiment analysis, outlining their approaches, advantages, and limitations, (3) summarize the application scenarios of LLM-based multimodal sentiment analysis technology, and (4) explore the challenges and potential research directions for multimodal sentiment analysis in the future.

Keywords text-centric, multimodal sentiment analysis, large language models, survey

Citation Yang H, Zhao Y Y, Wu Y, et al. Large language models meet text-centric multimodal sentiment analysis: a survey. Sci China Inf Sci, 2025, 68(10): 200101, https://doi.org/10.1007/s11432-024-4593-8

1 Introduction

Text-based sentiment analysis is a crucial research task in the field of natural language processing, aiming at automatically uncovering the underlying attitude that we hold towards textual content. However, humans often process emotions in a multi-modal environment, which differs from text-based scenarios in the following ways.

(1) Humans have the ability to acquire and integrate multimodal fine-grained signals. Humans often find themselves in multimodal scenarios, manifested as seamlessly understanding others' intentions and emotions through the combined effects of language, images, sound, and physiological signals. When processing emotions, humans have the ability to sensitively capture and integrate fine-grained sentiment signals from multiple modalities, and correlate them for emotional reasoning.

(2) Multimodal expression ability. The ways in which humans express emotions include language, facial expressions, body movements, and speech. For example, in daily conversations, our natural language expressions may be vague (such as someone saying "okay"), but when combined with other modal information, like visual modalities (e.g., a happy facial expression) or audio modalities (e.g., a prolonged intonation), the emotions expressed are different.

It is evident that the study of sentiment analysis within a multimodal context brings us closer to authentic human emotion processing. Research into multimodal sentiment analysis technologies [1,2] with human-like emotion processing capabilities will provide technical support for real-world applications such

* Corresponding author (email: yyzhao@ir.hit.edu.cn, siwei@iflytek.com)

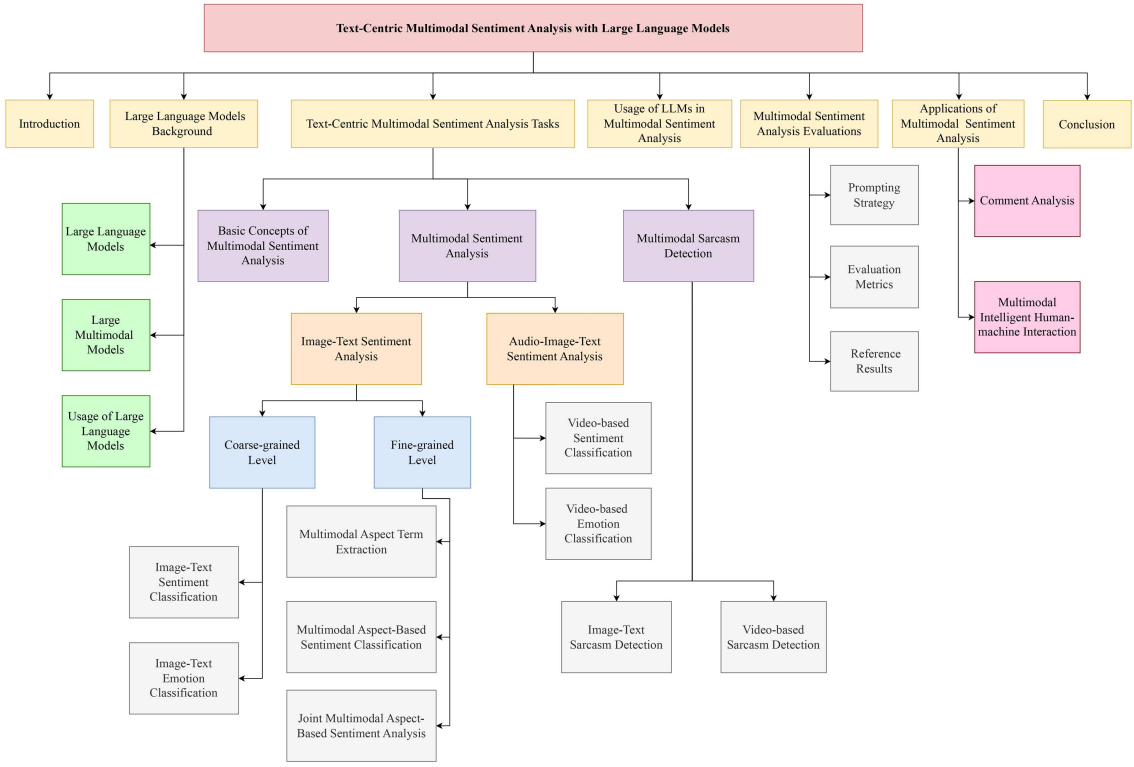

Figure 1 (Color online) Organization of the review article.

as high-quality intelligent companions, customer service, e-commerce, and depression detection. In recent years, large language models (LLMs) [3–5] have demonstrated astonishing human-machine conversational capabilities and showcased impressive performance across a wide range of natural language processing tasks, indicating their rich knowledge and powerful reasoning abilities. At the same time, large multimodal models (LMM) that increase the ability to understand modalities such as images also provide new ideas for multimodal-related tasks. They can directly perform tasks with zero-shot or few-shot context learning, requiring no supervised training [6–9]. While there have been some attempts to apply LLMs in text-based sentiment analysis [6, 10–13], there is a lack of systematic and comprehensive analysis regarding the application of LLMs and LMMs in multimodal sentiment analysis. Therefore, it remains unclear to what extent existing LLMs and LMMs can be used for multimodal sentiment analysis.

Given the crucial role of natural language in multi-modal sentiment analysis and its essential input for current LLMs and LMMs, we concentrate on text-centric multimodal sentiment analysis tasks that can leverage LLMs to enhance performance, such as image-text sentiment classification, image-text emotion classification, and audio-image-text (video) sentiment classification. In this work, we aim to provide a comprehensive review of the current state of text-centric multimodal sentiment analysis methods based on LLMs and LMMs. Specifically, we focus on the following questions. (1) How do LLMs and LMMs perform in a variety of multimodal sentiment analysis tasks? (2) What are the differences among approaches to utilize LLMs and LMMs in various multimodal sentiment analysis tasks, and what are their respective strengths and limitations? (3) What are the future application scenarios of multimodal sentiment analysis?

To this end, we first introduce the tasks and the most recent advancements in text-centric multimodal sentiment analysis. We also outline the primary challenges faced by current technologies and propose potential solutions. We examine a total of 14 multimodal sentiment analysis tasks, which have traditionally been studied independently. We analyze the distinct characteristics and commonalities of each task. The structure of the review study is depicted in Figure 1. Since LMMs are also based on LLMs, for convenience of presentation, the methods based on LLMs below include methods based on LMMs.

The rest of the sections of this paper are organized as follows. Section 2 introduces the background of LLMs and LMMs. In Section 3, we conduct an extensive survey on a wide range of text-centric multimodal sentiment analysis tasks, detailing the task definitions, related datasets and the latest methods. We also

summarize the advantages and advancements of LLM compared to previous techniques in multimodal sentiment analysis tasks in Section 4, as well as the challenges still faced. In Section 5, we introduced the prompt settings, evaluation metrics, and reference results related to LLM-based text-centric multimodal sentiment analysis methods. In Section 6, we look forward to the future application scenarios of multimodal sentiment analysis, followed by concluding remarks in Section 7.

2 Large language models background

2.1 Large language models

Generally, LLMs refer to transformer models with hundreds of billions (or more) of parameters, which are trained on large amounts of text data at a high cost, such as GPT-3 [3], PaLM [14], Galactica [15], and LLaMA2 [16]. LLMs typically possess extensive knowledge and demonstrate strong abilities in understanding, generating natural language, and solving complex tasks in practice. LLMs exhibit some abilities that are not present in small models, which is the most prominent feature that distinguishes LLM from previous pre-trained language models (PLMs), for example, in-context learning (ICL) capacity. Assuming that the language model has been provided with natural language instructions and several task demonstrations, it can generate the expected output of the test instance by completing the word sequence of the input text without additional training or gradient updates. *Instruction following.* By fine-tuning the mixture of multi-task datasets formatted through natural language descriptions (known as instruction adaptation), LLM performs well on unseen tasks also described in instruction form. Through fine-tuning instructions, LLM is able to follow task instructions for new tasks without using explicit examples, thus improving generalization ability. *Step-by-step reasoning.* For small language models (SLMs), it is often difficult to solve complex tasks involving multiple reasoning steps, such as mathematical word problems. Instead, using the chain-of-thought (CoT) cueing strategy [17–19], LLMs can solve such tasks by leveraging a cueing mechanism that involves intermediate reasoning steps to derive the final answer.

There have been some preliminary attempts to evaluate LLMs for text sentiment analysis tasks. In [6], the authors observed that the zero-shot performance of LLMs can be compared with fine-tuning BERT models [20]. In addition, in [11], the authors conducted preliminary research on some sentiment analysis tasks using ChatGPT, specifically studying its ability to handle polarity changes, open-domain scenarios, and emotional reasoning problems. In [12], the authors comprehensively tested the effectiveness of LLMs in text sentiment analysis datasets. In [21], the authors tested the effectiveness of commercial LLMs on a multimodal video-based sentiment analysis dataset. Despite these existing efforts, their scope is often limited to partial tasks and involves different datasets and experimental designs. Our goal is to comprehensively summarize the performance of LLMs in the field of multimodal sentiment analysis.

2.2 Large multimodal models

Large multimodal models (LMMs) are created to handle and integrate various data types, such as text, images, audio, and video. LMMs extend the capabilities of LLMs by incorporating additional modalities, allowing for a more comprehensive understanding and generation of diverse content. The development of LMMs is driven by the need to more accurately reflect the multimodal nature of human communication and perception. While traditional LLMs like GPT-4 are primarily text-based, LMMs are capable of processing and generating outputs across various data types. For instance, they can interpret visual inputs, generate textual descriptions from images, and even handle audio data, thus bridging the gap between different forms of information. One of the critical advancements in LMMs is the ability to create a unified multimodal embedding space. This involves using separate encoders for each modality to generate data-specific representations, which are then aligned into a cohesive multimodal space. This unified approach allows the models to integrate and correlate information from different sources seamlessly.

Notable examples include Gemini [22], GPT-4V, and ImageBind [23]. These models showcase the ability to process text, images, audio, and video, enhancing functionalities such as translation, image recognition, and more. In addition to these well-known models, other emerging models are also making significant strides: BLIP-2 [24] introduces a novel approach to integrate a frozen pre-trained visual encoder with a frozen large language model using a Q-former module. This module employs learnable input queries that interact with image features and the LLM, allowing for effective cross-modal learning. This setup helps maintain the versatility of the LLM while incorporating visual information effectively. LLaVA [25]

is a represent large multimodal model integrating a pre-trained CLIP [26] visual encoder (ViT-L/14), the Vicuna [27] language model, and a simple linear projection layer. Its training involves two stages: feature alignment pre-training, where only the projection layer is trained using 595k image-text pairs from Conceptual Captions dataset [28], and end-to-end fine-tuning, where the projection layer and LLM are fine-tuned using 158k instruction-following data and the ScienceQA dataset [29]. This setup ensures effective integration of visual and textual information, enabling LLaVA to excel in image captioning, visual question answering, and visual reasoning tasks. Qwen-VL [30] is a strong performer in the multimodal domain. Qwen-VL excels in tasks such as zero-shot image captioning and visual question answering, supporting both English and Chinese text recognition. Qwen-VL-Chat enhances interaction capabilities with multi-image inputs and multi-round question answering, showcasing significant improvements in understanding and generating multimodal content.

2.3 Usage of large language models

In [31], the authors summarized two paradigms for utilizing LLMs. **Parameter-frozen** application directly applies the prompting approach on LLMs without the need for parameter tuning. This category includes zero-shot and few-shot learning, depending on whether the few-shot demonstrations are required. **Parameter-tuning** application refers to the need for tuning parameters of LLMs. This category includes both full-parameter and parameter-efficient tuning, depending on whether fine-tuning is required for all model parameters.

In zero-shot learning, LLMs leverage the instruction following capabilities to solve downstream tasks based on a given instruction prompt, which is defined as

$$P = \text{Prompt}(I), \tag{1}$$

where I and P denote the input and output of prompting, respectively.

Few-shot learning uses in-context learning capabilities to solve the downstream tasks imitating few-shot demonstrations. Formally, given some demonstrations E, the process of few-shot learning is defined as

$$P = \text{Prompt}(E, I). \tag{2}$$

In the full-parameter tuning approach, all parameters of the model M are fine-tuned on the training dataset D

$$\widehat{M} = \text{Fine-tune}(M|D), \tag{3}$$

where $\widehat{M}$ is the fine-tuned model with the updated parameters.

Parameter-efficient tuning (PET) involves adjusting a set of existing parameters or incorporating additional tunable parameters (like bottleneck adapter [32], low-rank adaptation (LoRA) [33], prefix-tuning [34], and QLoRA [35]) to efficiently adapt models for specific downstream tasks. Formally, parameter-efficient tuning first tunes a set of parameters W, denoting as

$$\widehat{W} = \text{Fine-tune}(W|D, M), \tag{4}$$

where $\widehat{W}$ stands for the trained parameters.

3 Text-centric multimodal sentiment analysis tasks

Text-centric multimodal sentiment analysis mainly includes image-text sentiment analysis and audio-image-text (video) sentiment analysis. Among them, according to different emotional annotations, the two most common tasks are sentiment classification tasks (such as the most common three label classification tasks of positive, neutral, and negative) and emotion classification tasks (including emotional labels such as happy, sad, and angry). Similar to text-based sentiment classification, text-centered multimodal sentiment analysis can also be categorized into coarse-grained multimodal sentiment analysis (e.g., sentence-level) and fine-grained multimodal sentiment analysis (e.g., aspect-level) based on the granularity of the opinion targets. Existing fine-grained multimodal sentiment analysis usually focuses on image-text

Table 1 Categorization and representative methods for text-centric multimodal sentiment analysis.

Category		Task	Datasets	Methods
Image-text	Coarse-grained	Image-text sentiment classification	MVSA [36], MEMOTION 2 [37], MSED [38]	[39–49]
		Image-text emotion classification	TumEoM [51], MEMOTION 2 [37], MSED [38]	[47–51]
		Image-text sarcasm detection	MMSD [52], MMSD2.0 [53]	[52–64]
	Fine-grained	Multimodal aspect term extraction	Twitter-15 [65], Twitter-17 [65]	[66–72]
		Multimodal aspect sentiment classification	Multi-ZOL [73], Twitter-15 [65], Twitter-17 [65]	[48, 49, 65, 73–77]
		Joint multimodal aspect-sentiment analysis	Twitter-15 [65], Twitter-17 [65]	[78–87]
Audio-image-text	Video-based sentiment classification		ICT-MMMO [88], CMU-MOSI [89], CMU-MOSEI [90], CMU-MOSEAS [91], CH-SIMS [92], CH-SIMS 2 [93], MELD [94]	[58, 95–114]
	Video-based emotion classification		MELD [94], IEMOCAP [115], CMU-MOSEI [90], M3ED [116], MER2023 [117], EMER [118], ER2024 [119]	[94, 108–113] [115–119]
	Video-based sarcasm detection		MUStARD [120]	[120–122]

pair data, and includes multimodal aspect term extraction (MATE), multimodal aspect-based sentiment classification (MASC), and joint multimodal aspect-sentiment analysis (JMASA). Additionally, multimodal sarcasm detection has also become a widely discussed task in recent years. Due to the need to analyze conflicts between different modalities of sentiment, it highlights the importance of non-text modalities in sentiment judgment in real-world scenarios. We will introduce these tasks in the following subsections, and summarize them in Table 1 [36–122].

3.1 Basic concepts of multimodal sentiment analysis

Multimodal sentiment analysis (MSA) differs from traditional text-based sentiment analysis in that it combines multiple modalities, such as images and speech, to enhance the accuracy of sentiment classification. The most common multimodal sentiment analysis scenarios include "image-text", "audio-image" and "audio-image-text" (video). For example, the sentence "That's great!" expresses a positive emotion when analyzed as text alone, but when combined with an eye-rolling expression and a sharp tone of voice, the overall sentiment is sarcastically negative. Additionally, multimodal scenarios can also extend to more modalities that can reflect human emotions, such as "physiological signals" (skin conductance, electromyography, blood pressure, electroencephalography, respiration, pulse, electrocardiogram, etc.). In the following chapters of this paper, we will primarily focus on key tasks and techniques for text-centric multimodal sentiment analysis in "image-text" and "audio-image-text" (video) scenarios that can leverage LLMs. Since the "physiological signals" modality is interdisciplinary, encompassing fields like neuroscience and psychology, and has wide-ranging application potential, we will also provide a brief overview of it.

Although multimodal data contains richer information, effectively integrating multimodal information is a key challenge in current multimodal sentiment analysis tasks. Unlike sentiment expression in text-only modalities, sentiment expression in a multimodal context has its own particularities. (1) Complexity of sentiment semantic representation. In multimodal scenarios, sentiment semantics are derived from the representations of each participating modality. However, each single modality can have various representation methods, making the selection of which representation to use and how to fuse the representations from multiple modalities complex. (2) Complementarity of sentiment elements. Due to the participation of other modalities, the textual modality often has shorter and less informative expressions. Fine-grained sentiment elements from other modalities can provide effective supplements. (3) Inconsistency in sentiment expression. There can be conflicts in sentiment expressions among different modalities in the same scenario, with irony being the most common example.

Therefore, the core of multimodal sentiment analysis includes independent representation of single-modal sentiment semantics and fusion of multimodal sentiment semantic representations.

Independent representation of multimodal semantics refers to encoding each type of modality data separately. The encoding for each modality may take different forms and may not exist in the same semantic space. With the development of deep learning, deep learning techniques have shown outstanding performance in fields such as natural language processing, computer vision, and speech recognition. One of the greatest advantages is that many deep learning models (such as convolutional neural network (CNN) [123]) and concepts can be used across these three research areas, significantly lowering the research threshold for researchers and breaking down the barriers to joint representation of multimodal semantics. Each modality can be represented as vector information through deep learning models, and simple vector concatenation and addition can achieve the most basic multimodal semantic fusion, which serves as the basis for completing other multimodal downstream tasks. Additionally, researchers have

Table 2 Datasets of the text-centric multimodal sentiment analysis task. We use 'Emotions' to indicate that the dataset includes emotional labels, for example, happy, surprise, sad, and angry, and numeric intervals to represent the sentiment scoring annotations of the dataset.

Dataset	Language	Source	Year	Size	Modalities	Labels
ICT-MMMO [88]	English	YouTube	2011	340	A+V+T	$[-2, 2]$
IEMOCAP [115]	English	Shows	2008	10039	A+V+T	Emotions
CMU-MOSI [89]	English	YouTube	2016	2199	A+V+T	Neg, Neu, Pos
CMU-MOSEI [90]	English	YouTube	2018	23453	A+V+T	Neg, Neu, Pos and emotions
MELD [94]	English	Movies, TVs	2019	1443	A+V+T	Neg, Neu, Pos and emotions
CH-SIMS [92]	Chinese	Movies, TVs	2020	2281	A+V+T	$[-1, 1]$
CH-SIMS 2 [93]	Chinese	Movies, TVs	2022	4406	A+V+T	$[-1, 1]$
M3ED [116]	Chinese	Movies, TVs	2022	24449	A+V+T	Emotions
MER2023 [117]	Chinese	Movies, TVs	2023	3784	A+V+T	Emotions
EMER [118]	Chinese	Movies, TVs	2023	100	A+V+T	Emotions, reasoning
MER2024 [119]	Chinese	Movies, TVs	2024	6199	A+V+T	Emotions
CMU-MOSEAS [91]	Spanish, Portuguese, German, French	YouTube	2021	40000	A+V+T	$[-3, 3]$, [0,3]
UR-FUNNY [125]	English	Speech video	2023	16514	A+V+T	Funny
TumEoM [51]	English	Tumblr	2020	195264	V+T	Emotions
MVSA [36]	English	Twitter	2021	19598	V+T	Neg, Neu, Pos
Multi-ZOL [73]	Chinese	ZOL.com	2019	5288	V+T	[1,10]
MEMOTION 2 [37]	English	Reddit, Facebook	2022	10000	V+T	Neg, Neu, Pos
MSED [38]	English	Getty Image, Flickr and Twitter	2022	9190	V+T	Neg, Neu, Pos and emotions
Twitter-2015 [65]	English	Twitter	2019	5338	V+T	Neg, Neu, Pos
Twitter-2017 [65]	English	Twitter	2019	5972	V+T	Neg, Neu, Pos
MMSD [52]	English	Twitter	2019	24635	V+T	Neg, Pos
MMSD2.0 [53]	English	Twitter	2023	24635	V+T	Neg, Pos
MUStARD [120]	English	Movies, TVs	2021	690	A+V+T	Neg, Pos

found that each modality's representation is an independent modality space representation, residing in different vector spaces. Although rigid concatenation and addition have shown some effects, their theoretical significance is hard to justify. Therefore, scholars have begun to think about how to unify multiple modality representations into the same semantic space. For example, CLIP [26] uses techniques like contrastive learning and pre-training to obtain unified representations of images and text. This unified representation of multimodal semantics is also referred to as multimodal semantic fusion.

The fusion of multimodal sentiment semantic representation typically includes feature layer fusion, algorithm layer fusion, and decision layer fusion [124]. (1) Feature layer fusion (early fusion). This refers to the straightforward method of feature concatenation directly after extracting features from each modality. (2) Algorithm layer fusion (model-level fusion). This refers to thoroughly integrating each modality within different algorithmic frameworks. For example, two modalities can undergo nonlinear transformations through their respective deep learning models to achieve more abstract representations, sharing the same loss function to achieve comprehensive modality fusion. (3) Decision layer fusion (late fusion). This refers to combining each modality's representations with specific classification tasks to obtain independent representations for each modality and then using these to make the final classification decision. These approaches aim to address how to eliminate conflicts between modalities and how to achieve information complementarity among them.

Table 2 [36–38,51–53,65,73,88–94,115–120,125] provides an overview of 23 widely used datasets related to text-centric multimodal sentiment analysis. Each dataset is summarized based on aspects such as its modality composition, scale, and annotation type. This summary offers a comprehensive reference for understanding the characteristics and applicability of existing multimodal sentiment analysis datasets.

3.2 Image-text sentiment analysis

3.2.1 *Coarse-grained level*

Image-text coarse-grained sentiment analysis primarily encompasses two tasks: emotion classification and sentiment classification. Given an image-text pair, the emotion classification task aims to identify

emotional labels such as happiness, sadness, and surprise, inspired by the text-based emotion classification task. Sentiment classification aims to identify the sentiment label, which usually includes three categories (positive, neutral, negative). Problem formalization is as follows.

Given a set of multimodal posts from social media, $P = \{(T_1, V_1), \ldots, (T_N, V_N)\}$, where T_i is the text modality and V_i is the corresponding visual information, N represents the number of posts. We need to learn the model $f : P \to L$ to classify each post (T_i, V_i) into the predefined categories L_i. For polarity classification, $L_i \in \{\text{Positive}, \text{Neutral}, \text{Negative}\}$; for emotion classification, $L_i \in \{\text{Angry}, \text{Bored}, \text{Calm}, \text{Fear}, \text{Happy}, \text{Love}, \text{Sad}\}$.

The earliest image-text sentiment classification models were feature-based. In [40], the authors used SentiBank to extract 1200 adjective-noun pairs (ANPs) as visual semantic features and employed SentiStrength [126] to compute text sentiment features for handling multimodal tweet sentiment analysis. In [41], the authors presented a cross-media bag-of-words model to represent the text and image of a Weibo tweet as a unified bag-of-words representation. Then some neural network models showed better performance. In [42, 43], the authors used CNN models to get the representation of text and image. In [44], the authors believed that more detailed semantic information in the image is important and constructed HSAN, a hierarchical semantic attentional network based on image caption for coarse-level multimodal sentiment analysis. MultiSentiNet [45] focused on the correlation between images and text, aggregating the representation of informative words with visual semantic features, objects, and scenes. Considering the mutual influence between image and text, Co-Mem [46] is designed to iteratively model the interactions between visual contents and textual words for multimodal sentiment analysis.

In [39], the authors found that images play a supporting role to text in many sentiment detection cases, and proposed VistaNet, which instead of using visual information as features only rely on visual information as alignment for pointing out the important sentences of a document using attention. With respect to each image representation f_j^v, the goal is to learn the attention weights $\beta_{j,i}$ for text representations f_j^t:

$$p_j = \tanh(W_p f_j^v + b_p), \tag{5}$$

$$q_i = \tanh(W_q f_i^t + b_q), \tag{6}$$

$$v_{j,i} = V^\top (p_j \odot q_i + q_i), \tag{7}$$

$$\beta_{j,i} = \frac{\exp(v_{j,i})}{\sum_i \exp(v_{j,i})}, \tag{8}$$

where W_p, W_q, b_p, b_q are learnable parameters, firstly, projecting both image representation and text representation onto an attention space followed by a non-linear activation function tanh. Then, let the image projection p_j interact with the sentence projection q_i in two ways: element-wise multiplication and summation. The learned vector V plays the role of global attention context.

CLMLF [127] applies contrastive learning and data augmentation to align and fuse the token-level features of text and image. In addition to focusing on sentiment, emotions are equally important. In [51], the authors built an image-text emotion dataset, named TumEmo, and further proposed MVAN for multi-modal emotion analysis. In [47], the authors observed that multimodal emotion expressions have specific global features and introduced a graph neural network, proposing an emotion-aware multichannel graph neural network method called MGNNS. MULSER [50] is also a graph-based fusion method that not only investigates the semantic relationship among objects and words respectively, but also explores the semantic relationship between regional objects and global concepts, which has also yielded effective results.

Traditional non-LLM multimodal sentiment analysis methods typically rely on feature fusion. However, these methods often use lightweight models for textual feature extraction, which lack the ability to capture deep contextual information and world knowledge. As a result, they face limitations in accurately interpreting the emotions conveyed by multimodal content such as images and text. The emergence of large models has significantly improved the understanding capabilities for the text modality and demonstrated stronger generalization abilities. Multimodal sentiment analysis approaches that leverage LLMs and LMMs have shown superior performance. For example, WisdoM [128] leverages the contextual world knowledge induced from the LMMs for enhanced multimodal sentiment analysis. The process involves three stages. (1) Prompt templates generation. Using ChatGPT to create templates that help LMMs understand the context better. (2) Context generation. Feeding these templates into LMMs along with the

sentence and image to generate rich contextual information. (3) Contextual fusion. Combining this contextual information with the original sentiment predictions to enhance accuracy, particularly for difficult samples. A training-free module called contextual fusion is introduced to minimize noise in the contextual data, ensuring that only relevant information is considered during sentiment analysis. WisdoM significantly outperforms existing state-of-the-art methods in MSED dataset, demonstrating its effectiveness in integrating contextual knowledge for improved sentiment classification. In addition, inspired by the success of textual prompt-based fine-tuning approaches in few-shot scenario, the authors [48] introduced a multi-modal prompt-based fine-tuning approach UP-MPF, and the authors [49] proposed a prompt-based vision-aware language modeling (PVLM) for multimodal sentiment analysis.

We summarize the commonly used datasets for coarse-grained image-text sentiment and emotion analysis, including TumEom, MVSA, MEMOTION 2, and MSED.

TumEmo is a multimodal weak-supervision emotion dataset containing a large amount of image-text data crawled from Tumblr. The dataset contains 195265 image-text pairs with 7 emotion labels: angry, bored, calm, fearful, happy, loving, and sad.

MVSA dataset is collected from image-text pairs on the Twitter platform and is manually annotated with three sentiment labels: positive, neutral, and negative. The MVSA dataset consists of two parts: MVSA-Single, where each sample is annotated by a single annotator, comprising 4869 image-text pairs, and MVSA-Multiple, where each sample is annotated by three annotators with three emotion labels, totaling 19598 image-text pairs. The MVSA corpus is another example of coarse-grained multimodal sentiment classification dataset.

MEMOTION 2 is a dataset focused on classifying emotions and their intensities into discrete labels. It includes 10000 memes collected from various social media sites. These memes are typically humorous and aim to evoke a response. Overall sentiment (positive, neutral, negative), emotion (humour, sarcasm, offence, motivation), and scale of emotion are all annotated for each meme (0–4 levels).

MSED comprises 9190 pairs of text and images sourced from diverse social media platforms, including but not limited to Twitter, Getty Images, and Flickr. Each piece of multi-modal sample is manually annotated with desire category, sentiment category (i.e., positive, neutral and negative) and emotion category (happiness, sad, neutral, disgust, anger and fear).

3.2.2 *Fine-grained level*

Image-text fine-grained sentiment analysis focuses on analyzing sentiment elements that are finer than sentence-level, such as aspect term (a) and sentiment polarity (p), or their combinations. It has received widespread attention in recent years and mainly includes three subtasks: multimodal aspect term extraction (MATE), multimodal aspect-based sentiment classification (MASC) and joint multimodal aspect-sentiment analysis (JMASA). We illustrate the definitions of all the sub-tasks with a specific example in Figure 2.

Multimodal aspect term extraction. As shown in Figure 2, MATE aims to extract all the aspect terms mentioned in a sentence. Given the multimodal input includes a n-words sentence $S = (w_1, w_2, \ldots, w_n)$ and a corresponding image I, the goal of MATE is to predict the label of each word in scheme $y_i \in \{B, I, O\}$, where B indicates the beginning, I indicates the inside and the end of an aspect term, O means non-target words.

Inspired by text-based aspect term extraction methods [129–131], MATE approaches usually view this task as a sequence labeling problem. How to utilize visual information to improve the accuracy of aspect term recognition is the key to this task. Some studies [67,71] focused on named entity recognition suggest using ResNet encoding to leverage whole image information to enhance the representation of each word. Various neural network-based methods have been developed, including those using recurrent neural networks [67, 68], Transformers [66, 69, 70], and graph neural networks [72]. Conditional random fields (CRF) are widely used in sequence labeling tasks because CRF considers the correlations between labels in neighborhoods. For example, an adjective has a greater probability of being followed by a noun than a verb in POS tagging task. Using $Y = (y_1, y_2, \ldots, y_n)$ represents a generic sequence of labels for input S. Given sequence S, all the possible label sequences Y can be calculated by the following equation:

$$p(Y|S) = \frac{\prod_{i=1}^{n} \Omega(y_{i-1}, y_i, X)}{\sum_{y' \in Y} \prod_{i=1}^{n} \Omega(y'_{i-1}, y'_i, X)}, \tag{9}$$

where $\Omega(y_{i-1}, y_i, X)$ and $\Omega(y'_{i-1}, y'_i, X)$ are potential functions.

Text: RT @ EarthPicturz : [Kilauea]$^{p_1:\,neutral}_{a_1}$, on the [Big Island of Hawaii]$^{p_2:\,neutral}_{a_2}$, sends streams of lava steaming into the Pacific Ocean.

Image:

Sub-tasks	Input	Output
Multimodal Aspect Term Extraction (MATE)	S+T	$\{a_1, a_2\}$
Multimodal Aspect-Oriented Sentiment Classification (MASC)	S+T+a_1	p_1
	S+T+a_2	p_2
Joint Multimodal Aspect-Sentiment Analysis (JMASA)	S+T	$(a_1, p_1), (a_2, p_2)$

Figure 2 (Color online) Image-text fine-grained sentiment analysis tasks. The case is sourced from the publicly available Twitter-2015 [65] dataset, which was collected and released solely for research purposes.

However, these methods are relatively independent and often focus more on entity information while neglecting the emotional information of the target. Therefore, as research progresses, more scholars in multimodal scenarios are not only extracting aspect terms but also jointly performing corresponding sentiment classification.

For example, inspired by the fuzzy span universal information extraction (FSUIE) framework [132], which emphasizes the representation of short span-length features, DQPSA [133] formulates MATE and MASC as span recognition tasks within a unified framework, thereby avoiding the complexity of traditional sequence generation approaches. Rather than relying solely on exact boundary labels, FSUIE introduces a fuzzy span mechanism, in which span boundaries are represented as probability distributions instead of fixed positions. Under this formulation, the model produces a distribution $\hat{q}$ over potential boundary positions (termed fuzzy boundaries), capturing uncertainty and contextual ambiguity.

The fuzzy span loss is defined as the Kullback-Leibler divergence between the predicted distribution p and the fuzzy gold distribution $\hat{q}$, and is integrated into the overall training objective:

$$L_{\mathrm{FS}} = D_{\mathrm{KL}}(\hat{q}\|p) = \sum_{i=1}^{N} \hat{q}(x_i) \log \frac{\hat{q}(x_i)}{p(x_i)}, \tag{10}$$

where p denotes the model's predicted boundary distribution and $\hat{q}$ is derived from annotations via the fuzzy span distribution generator (FSDG). The total loss function combines the original binary cross-entropy (BCE) loss L_{ori} with the fuzzy span loss, scaled by a weighting coefficient λ:

$$L = L_{\mathrm{ori}} + \lambda L_{\mathrm{FS}}. \tag{11}$$

This approach enables the model to better capture soft boundary information by treating span boundaries as probabilistic regions rather than fixed points.

As described in [133], the DQPSA framework comprises four main components: a frozen image encoder, a prompt-as-dual-query module, a text encoder, and an energy-based pairwise expert. To enhance alignment between visual input and the analysis target, the prompt-as dual-query module allows prompts to interact with both image and text modalities, facilitating more accurate information extraction through visual-textual fusion.

Furthermore, to capture the semantic dependency between start and end positions within a span, DQPSA leverages an energy-based model [134] implemented via the energy-based pairwise expert. Rather than predicting boundary positions independently, this module evaluates spans through pairwise scoring, measuring the stability and compatibility of boundary pairs. By incorporating global span-level interactions, this design leads to more coherent and accurate span prediction.

Multimodal aspect-based sentiment classification. As shown in Figure 2, MASC aims to identify the sentiment polarity of a given aspect term in a sentence. Problem formalization as follows: given a set of multimodal samples $S = \{X_1, X_2, \ldots, X_{|S|}\}$, where $|S|$ is the number of samples. And for each sample, we are given an image $V \in \mathbb{R}^{3\times H\times W}$ where 3, H and W represent the number of channels, height and width of the image, and an N-word textual content $T = (w_1, w_2, \ldots, w_N)$ which contains an M-word sub-sequence as target aspect $A = (w_1, w_2, \ldots, w_M)$. The goal of MASC is to learn a sentiment classifier to predict a sentiment label $y \in \{\text{Positive}, \text{Negative}, \text{Neutral}\}$ for each sample $X = (V, T, A)$.

Different from text-based aspect sentiment classification [135–137], it is challenging to effectively discover visual sentiment information and fuse it with textual sentiment information. In [73], the authors constructed the Multi-ZOL dataset for the MASC task. This dataset collects and organizes comments about smartphones from the ZOL.com business portal website. At the same time, they proposed a multimodal interactive memory network (MIMN) based on an attention mechanism to capture the information interaction between different modalities. In addition, other researchers [65, 75] have proposed models like the LSTM-based ESAFN model and the Transformer-based TomBERT model for the MASC task, enhancing the interaction of inter-modal and intra-modal sentiment information is the core of these models. The TomBERT model treats the hidden states of the target aspect A as queries, and the regional image features H^V as keys and values, such that the target is leveraged to guide the model to align it with the appropriate regions.

$$\text{ATT}_i(A, H^V) = \text{softmax}\left(\frac{[W_{Q_i}A]^T[W_{K_i}H^V]}{\sqrt{d//m}}\right)[W_{V_i}H^V]^T, \tag{12}$$

where $W_{Q_i}, W_{K_i}, W_{V_i} \in R^{d/m\times d}$ are parameters, m represent the attention head number, d is the input embeddings dimension.

Compared with other multimodal tasks such as image and text retrieval, the sentiment annotation used in the MASC task lack strong supervision signals for cross-modal alignment. This issue makes it difficult for MASC models to learn cross-modal interactions and causes models to learn the bias brought by the image. In [74], the authors proposed a new method to utilize visual modalities, the image caption generation module in their model undertakes the task of cross-modal alignment. They convert images into text descriptions based on the idea of cross-modal translation.

$$C = \text{Caption\_Transformer}(V), \tag{13}$$

where C and Caption_transformer denote the output image captions and the transformer-based image caption generator, respectively. In sentence-pair classification mode, input to pre-trained language model takes the sentence-pair form

$$[\text{CLS}]w_1^T, w_2^T, \ldots, w_{T.\text{len}}^T[\text{SEP}]w_1^C, w_2^C, \ldots, w_{C.\text{len}}^C[\text{PAD}], \ldots, [\text{PAD}], \tag{14}$$

where w_i^T are the tokens of the input text, and w_i^C are the tokens of the image caption. In [76], the authors continued with the idea of modal transformation and employed facial emotions as a supervised signal for learning visual emotions.

The pre-trained model SMP [138] employs a pre-training task that considers fine-grained emotional information. Compared to previous studies that only used single-modal models or fine-tuned multimodal models without considering emotional information during the pre-training phase, it achieves significant performance improvements.

As large language models evolve, LLMs and LMMs have been adapted to various tasks [17, 139–141], yet their application to the MASC task remains at an early stage. In [77], the authors proposed A^{2}II, a multimodal sentiment analysis model based on instruction tuning. A^{2}II takes a sentence, an associated image, and an aspect term as input, and aims to predict the sentiment polarity toward the given aspect.

To improve the alignment between visual content and textual aspect, the model first encodes the image using a visual encoder to obtain global visual features. Then, a Q-Former module is employed to align

these features with the aspect term, producing a unified multimodal representation. Recognizing that the image may not always be relevant to the aspect, A^2II introduces a plug-and-play instruction selector that dynamically chooses the most appropriate instruction from a predefined instruction pool based on the multimodal context. This selected instruction is combined with the input sentence to form an enriched textual prompt.

Finally, the model fuses the multimodal representation with the instruction-augmented text and feeds the combined input into a language model to generate the sentiment prediction. This framework enhances robustness by filtering out irrelevant visual information and tailoring the instruction to the input context, thereby improving performance in aspect-level multimodal sentiment classification.

Joint multimodal aspect-sentiment analysis. As shown in Figure 2, JMASA aims to extract all aspect terms and their corresponding sentiment polarities simultaneously. Problem formalization as follows: given a collection of multimodal sentence-image pairs, denoted as M. Each pair $m_i \in M$ comprises a sentence $Si = (w_1, w_2, \ldots, w_n)$ and a corresponding image v_i. The objective of JMASA is to predict the corresponding aspect-sentiment pair $y = (y_1, y_2, \ldots, y_n)$ for each sentence-image pair. Here, $y_i \in \{B\text{-POS}, I\text{-POS}, B\text{-NEG}, I\text{-NEG}, B\text{-NEU}, I\text{-NEU}\} \cup O$. In this case, B refers to the initial token of the aspect term, I indicates tokens within the specific aspect term and O indicates words "outside" the specific aspect. Moreover, POS, NEU, and NEG are abbreviations for positive, neutral, and negative sentiments associated with the specific aspect.

As a pioneer, in [79], the authors proposed joint multimodal aspect-sentiment analysis, which jointly performs multimodal aspect term extraction and multimodal aspect sentiment classification. Since it is a joint task with aspect terms extraction and aspect sentiment classification, the authors calculate two different sets of loss simultaneously as follows:

$$L = -\sum_{i=1}^{k} y_i^s \log p_i^{\text{str}} - \sum_{i=1}^{k} y_i^e \log p_i^{\text{end}} - \sum_{t=1}^{m} \sum_{i=1}^{\epsilon} y_{ti}^p \log p_{ti}^p, \tag{15}$$

where y^s, y^e, y^p are one-hot labels indicating golden start, end positions, true sentiment polarity separately, and a, m are the number of sentence tokens, aspects respectively.

Benefiting from the advancements in visual-language pre-train models, in [78], the authors have designed multimodal sentiment pre-training tasks and developed a unified multimodal encoder-decoder architecture pre-training model for JMASA. In [80], the authors utilized a cross-modal multi-task transformer (CMMT) to derive sentiment-aware features for each modality and dynamically control the impact of visual information on textual content during inter-modal interaction. However, the innate semantic gap between visual and language modalities remains a huge challenge for the use of these methods, in [81], the authors believed that the aesthetic attributes of images potentially convey a more profound emotional expression than basic image features and proposed Atlantis. Some scholars [82] have also noticed the impact of image-text pair quality, finding that many studies have overestimated the importance of images due to the presence of many noise images unrelated to text in the datasets. Drawing from the concept of curriculum learning, they proposed a multi-grained multi-curriculum denoising framework (M2DF), which achieves denoising by adjusting the order of the training data. AOM [83] is designed with an aspect-aware attention module that simultaneously selects text tokens and image blocks semantically related to the aspect to detect semantic and emotional information related to the aspect, thereby reducing noise introduced during the cross-modal alignment process. RNG [84] to simultaneously reduce multi-level modality noise and multi-grained semantic gap, design three constraints: (i) global relevance constraint (GR-Con) based on text-image similarity for instance-level noise reduction, (ii) information bottleneck constraint (IB-Con) based on the information bottleneck (IB) principle for feature-level noise reduction, and (iii) semantic consistency constraint (SC-Con) based on mutual information maximization in a contrastive learning way for multi-grained semantic gap reduction. To bridge the semantic gap between modal spaces and address the interference of irrelevant visual objects at different scales, in [85], the authors proposed a multi-level text-visual alignment and fusion network (MTVAF).

With the help of LLMs, the JMASA task has also seen further development in recent years. In [86], the authors observed that converting MASC into a masked language modeling (MLM) task, as done in PVLM [49] and UP-MPF [48], was not well suited to JMASA and MATE tasks. In response, they proposed the generative multimodal prompt (GMP) model. More recently, Ref. [87] explored the use of ChatGPT for in-context learning (ICL) on the JMASA task and proposed a versatile ICL framework to support both zero-shot and few-shot learning.

Specifically, the framework includes three key modules: input construction, in-context learning, and a demonstration exemplar retriever. To prepare inputs, the model first transforms visual content into textual form by generating image captions and extracting visual entities, entity types, and visual sentiments. These visual texts are then combined with the original input text. For few-shot scenarios, the framework employs an entity-aware contrastive learning model to retrieve top-K training samples most similar to the test instance. This retriever is trained using a scoring function that computes semantic similarity among samples, allowing for the creation of positive and negative instance pairs used in contrastive learning.

Once the most relevant samples are retrieved, they are used as demonstrations in the ICL prompts, improving the alignment between the current input and the contextual examples. Compared to random demonstration selection, this entity-aware method leads to more effective prompt construction and ultimately improves performance within multimodal sentiment analysis.

We summarize the commonly used datasets for fine-grained image-text sentiment analysis, including Multi-ZOL, Twitter-15 and Twitter-17.

Multi-ZOL collects and organizes comments about smartphones from the ZOL.com business portal website. Multi-ZOL dataset includes sentiment ratings for six aspects, such as price-performance ratio, performance configuration, battery life, appearance and feeling, photographing effect, and screen. For each aspect, the comment has an integer sentiment score from 1 to 10, which is used as the sentiment label.

Twitter-2015 and **Twitter-2017** datasets are commonly used datasets for fine-grained image-text sentiment analysis tasks. These datasets are collected from English tweets on the social media platform Twitter and are in the form of image-text pairs. The datasets provide annotations for aspects mentioned in the text. Sentiment labels are categorized into three classes: positive, neutral and negative. Specifically, the Twitter-2015 dataset contains 5338 tweets with images, while the Twitter-2017 dataset contains 5972 tweets with images.

However, existing benchmark datasets used in the MASC task, such as Twitter-2015 and Twitter-2017, have shown limitations in supporting true multimodal learning. Recent work by Ye et al. [142] conducted a comprehensive empirical study and analysis of these datasets. Their findings reveal that the sentiment polarity of most targets can be determined solely by the text, rendering the visual modality less informative. Furthermore, a large portion of images in these datasets either lack the target object or provide noisy, irrelevant visual content. As a result, multimodal models fail to significantly outperform strong text-only baselines. The authors highlight the urgent need for better-curated datasets where visual information contributes essential complementary sentiment cues to the textual context.

3.3 Audio-image-text sentiment analysis

Audio-image-text (video) sentiment analysis differs from image-text sentiment analysis in two main aspects. (1) Different data emphasis. Existing text-image sentiment datasets are drawn from social media and e-commerce platforms, covering a wide range of content. In contrast, visual information in video sentiment datasets often focuses on the facial expressions and body movements of speakers. (2) Videos can be considered as temporal sequences of text-image pairs, necessitating considerations of intra-modal emotional factors in audio sequences and video frame sequences, as well as alignment relationships between text, video frames, and audio over time. Video-based sentiment analysis primarily includes sentiment classification and emotion classification tasks. Sentiment classification involves three, five, or seven-category classification tasks, while emotion classification comprises multi-label emotion recognition (where each sample corresponds to multiple emotion labels) and single-label emotion recognition. Common emotion labels include happiness, surprise, and anger. Problem formalization as follows.

In audio-image-text sentiment analysis tasks, the input is utterance consisting of three modalities: textual, acoustic and visual modality, where $m \in \{t, a, v\}$. The sequences of these three modalities are represented as triplet (T, A, V), including $T \in \mathbb{R}^{N_t \times d_t}$, $A \in \mathbb{R}^{N_a \times d_a}$ and $V \in \mathbb{R}^{N_v \times d_v}$ where N_m denotes the sequence length of corresponding modality and d_m denotes the dimensionality. The goal of audio-image-text sentiment analysis tasks is to learn a mapping $f(T, A, V)$ to infer the sentiment score $\hat{y} \in \mathbb{R}$.

As the audio-image-text sentiment analysis methods proposed by scholars in recent years generally cater to both sentiment classification and emotion classification tasks, this paper will review the existing multimodal sentiment analysis methods around two core themes: cross-modal sentiment semantic alignment and multimodal sentiment semantic fusion.

3.3.1 *Cross-modal sentiment semantic alignment*

Cross-modal sentiment semantic alignment methods aim to explore the associations between emotional information across different modalities, analyze the corresponding relationships between them (alignment relationship modeling), and reduce the semantic distance between representations across modalities (semantic representation alignment). Cross-modal sentiment semantic alignment methods can help overcome the challenges brought by the semantic gap of heterogeneous modalities and are a prerequisite for multimodal sentiment semantic fusion methods. Specifically, by exploring the alignment relationships between different modal sentient semantic representations, these methods can help the fusion model ignore irrelevant information and focus on modeling effective information. By bringing emotional semantic representations closer in the representation space, these methods can reduce modal differences between representations, lower the difficulty of fusion, and increase fusion efficiency. This paper surveys existing cross-modal sentiment semantic alignment methods and categorizes them into three types based on different alignment strategies and purposes: attention-based alignment, contrastive learning-based alignment, and cross-domain transfer learning-based alignment.

Attention-based alignment. The attention mechanism has been proven to be an effective method for cross-modal semantic alignment in the field of multimodal learning [143]. Not only can the attention mechanism learn to adapt the alignment relationships for specific tasks through the optimization of task-specific objective functions, but it can also provide a degree of interpretability by outputting attention weights. For example, in the field of image captioning, the attention mechanism focuses on relevant areas when generating text words [144], demonstrating the alignment relationship between words and image regions. Inspired by related research in the multimodal learning field, in [104], the authors proposed using a cross-modal attention mechanism to learn the alignment relationships between pairs of modalities and developed a transformer-based multimodal sentiment analysis model named MulT. The core of the MulT model lies in modeling cross-modal alignment relationships by inserting cross-modal attention layers into the transformer module, allowing dynamic alignment and fusion of fine-grained sentiment information from various modalities. First, use α and β to represent two different modalities, define the Querys as $Q_\alpha = X_\alpha W_{Q_\alpha}$, Keys as $K_\beta = X_\beta W_{K_\beta}$, and Values as $V_\beta = X_\beta W_{V_\beta}$, where $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}, W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$ and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ are weights. The latent adaptation from β to α is presented as the crossmodal attention $Y_\alpha := \mathrm{CM}_{\beta\to\alpha}(X_\alpha, X_B) \in \mathbb{R}^{T_\alpha \times d_v}$:

$$Y_\alpha = \mathrm{CM}_{\beta\to\alpha}(X_\alpha, X_\beta) = \mathrm{softmax}\left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\right) V_\beta = \mathrm{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}\right) X_\beta W_{V_\beta}. \tag{16}$$

Note that Y_α has the same length as Q_α, but is meanwhile represented in the feature space of V_β. Specifically, the scaled (by $\sqrt{d_k}$) softmax computes a score matrix $\mathrm{softmax}(\cdot) \in \mathbb{R}^{T_\alpha \times T_\beta}$, whose (i, j)-th entry measures the attention given by the i-th time step of modality α to the j-th time step of modality β. Hence, the i-th time step of Y_α is a weighted summary of V_β, with the weight determined by i-th row in $\mathrm{softmax}(\cdot)$.

Building on the cross-modal attention mechanism designed in MulT, in [105], the authors introduced the cubic attention mechanism, which generates a three-dimensional attention tensor through parameter computations, representing the alignment information among the three modal representations.

Contrastive learning-based alignment. Contrastive learning achieves cross-modal representation alignment by bringing the representations of positive examples closer together and pushing the representations of negative examples farther apart. A classic model in the field of multimodal learning, CLIP [26], uses contrastive learning to align the semantic representations of text and image modalities, significantly enhancing the quality of image representations and achieving excellent results in tasks such as zero-shot image classification. Inspired by this, the field of audio-image-text sentiment analysis has adopted contrastive learning methods for sentiment semantic representation alignment. In [106], the authors proposed achieving cross-modal emotional semantic alignment by bringing closer the representations of different modalities within the same sample. In [107], the authors suggested using the text-audio and text-image modal information of input samples to predict the corresponding image and audio representations of the samples, then aligning the predicted representations with the actual ones and distancing representations from different samples, thereby aligning the semantic representations of different modalities within the same sample. The proposed model contains two key modules, the uni-modal coding drive the model to focus on informative features, which then implicitly filter out inherent noise and produce robust and effective uni-modal representation for acoustic and visual modalities.

Given a batch set $F_{uni} = \{F_u^0, F_u^1, \ldots, F_u^{n-1}\}$, noted that there is a single positive key $F_u^\dagger$ (as k^+) that each encoded query F_u^i (as q, $i \in [1, n]$) matches, while the other representations F_u^j ($j \in [0, n]$ and $j \neq i$) in the same batch are considered as negative key samples k^-. With the similarity measured by dot product, the uni-modal instance contrastive loss $\mathcal{L}_{uni}$ in

$$\mathcal{L}_{uni}^u \triangleq -\log \frac{\exp(q \cdot k^+/\tau)}{\sum_{i=1}^n \exp(q \cdot q^i/\tau)} = -\mathop{\mathbb{E}}_{F_{uni}} \left[\log \frac{\exp(F_u \cdot F_u^\dagger/\tau)}{\sum_{i=1}^n \exp(F_u \cdot F_u^i/\tau)}\right], \tag{17}$$

where τ is a temperature hyper-parameter that controls the probability distribution over distinct instances. Due to $u \in \{a, v\}$, the final uni-modal instance contrastive loss $\mathcal{L}_{uni} = \mathcal{L}_{uni}^a + \mathcal{L}_{uni}^v$.

The cross-modal prediction captures commonalities among different modalities and outputs predictive representation full of interaction dynamics. Each query q has a corresponding key as k^+ while the other representations in the same batch are seen as k^-. Similar with uni-modal instance contrastive loss $\mathcal{L}_{uni}$, the cross-modal instance contrastive loss $\mathcal{L}_{\text{cross}}$ is presented as

$$\mathcal{L}_{\text{cross}} \triangleq -\mathop{\mathbb{E}}_{F_{\text{cross}}} \left[\log \frac{\exp(F_c \cdot F_c^+/\tau)}{\sum_{i=1}^n \exp(F_c \cdot F_c^i/\tau)}\right], \tag{18}$$

where $F_c \backslash F_c^+ \in \{P_u, G_u\}$, P_u represent the prediction while G_u represent the target, $u \in \{a, v\}$ and $F_{\text{cross}} = \{F_c^1, \ldots, F_c^n\}$.

Cross-domain transfer learning-based alignment. The field of cross-domain transfer learning primarily studies how to align the sample spaces of target domains with those of source domains so that classifiers trained in the source domains can be directly reused in the target domains. The objectives of this field align broadly with those of cross-modal sentiment representation alignment, hence some studies have explored using cross-domain transfer learning methods for sentiment semantic representation alignment. In [108], considering the rich information content of textual representations, the authors proposed using deep canonical correlation analysis (DCCA) to align audio and visual representations with textual representations, thereby enhancing the audio and visual representations. In [97], the authors explored using a metric-based domain transfer method, utilizing central moment discrepancy (CMD) to design a loss function that aligns the representations of the three modalities within the same sample. The overall learning of the model is performed by minimizing

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\, \mathcal{L}_{\text{sim}} + \beta\, \mathcal{L}_{\text{diff}} + \gamma\, \mathcal{L}_{\text{recon}}, \tag{19}$$

where α, β, γ are the interaction weights that determine the contribution of each regularization component to the overall loss $\mathcal{L}$. Minimizing the similarity loss $\mathcal{L}_{\text{sim}}$ reduces the discrepancy between the shared representations of each modality. This helps the common cross-modal features to be aligned together in the shared subspace. Difference loss $\mathcal{L}_{\text{diff}}$ is to ensure that the modality-invariant and -specific representations capture different aspects of the input. As the difference loss is enforced, there remains a risk of learning trivial representations by the modality-specific encoders. To avoid this situation, the authors add a reconstruction loss $\mathcal{L}_{\text{recon}}$ that ensures the hidden representations to capture details of their respective modality. The task-specific loss $\mathcal{L}_{\text{task}}$ estimates the quality of prediction during training.

In [109], the authors have employed adversarial learning methods to align sentiment semantic representations across different modalities.

3.3.2 *Multimodal sentiment semantic fusion*

Multimodal sentiment semantic fusion aims to efficiently aggregate sentiment information from different modalities to achieve comprehensive and accurate sentiment understanding. The challenge of fusion lies in how to fully capture the complex interactions among multimodal sentiment semantic information, thereby facilitating sentiment reasoning and prediction. This paper surveys existing multimodal sentiment semantic fusion methods and categorizes them into three types: tensor-based fusion, fine-grained temporal interaction modeling fusion, and pre-trained model-based fusion.

Tensor-based fusion. In the early stages of audio-image-text sentiment analysis research, considering the small scale of datasets and limited computational resources, researchers represented the raw inputs of each modality as a single emotional semantic representation before proceeding to multimodal emotional semantic representation fusion. The simplest fusion strategy was to directly concatenate the emotional semantic representations of different modalities, but this method did not explicitly model the

higher-order interactions between emotional information from different modalities. To address this issue, in [95], the authors proposed using the outer product of vectors to fuse different modal representations, thereby modeling interactions among unimodal, bimodal, and trimodal emotional semantic representations simultaneously. However, this method, due to the complexity of the outer product operation being tied to the product of input vector dimensions, resulted in high computational costs and slow efficiency.

To improve computational efficiency, the authors in [110] proposed the low-rank multimodal fusion (LMF) method. Instead of directly learning the high-dimensional weight tensor used for modality fusion, LMF factorizes this tensor into a sum of outer products of low-rank modality-specific matrices. Specifically, the high-order weight tensor is decomposed into several sets of vectors, each corresponding to a modality, and parameterized with a fixed number of decomposition factors (i.e., rank-r factors). These modality-specific low-rank factors are then combined using outer products to reconstruct a low-rank approximation of the original fusion tensor. During forward computation, the fusion result is obtained by applying this approximated tensor to the modality features. This approach significantly reduces the number of parameters and computational cost, while still modeling multimodal interactions.

In [111], the authors introduced a three-stage multimodal emotional representation fusion strategy consisting of representation slice grouping, intra-group representation slice fusion, and global representation fusion. Representation slice grouping involves splitting the representations of each modality into the same fixed number of small groups, numbering them, and then locally fusing representation slices of the same number from different modalities together. This approach reduces the dimensions of representations to be fused later, thereby enhancing fusion efficiency. Intra-group representation slice fusion uses the outer product method to fuse the representation slices of the three modalities within the group, which, due to the smaller feature dimensions, significantly speeds up the fusion process. Finally, long short-term memory (LSTM) networks are used to perform global representation fusion of the different groups after fusion. This method reduces the computational complexity of the tensor outer product fusion method to some extent through block processing.

Fine-grained temporal interaction modeling fusion. This type of fusion method focuses on capturing more localized, fine-grained interactions of multimodal information. These methods first obtain fine-grained representations corresponding to each time step of each modality, and then perform multimodal sentiment semantic fusion based on these representations to capture the interactions between cross-modal and cross-temporal sentiment information. In [96], the RAVEN model is a typical method in this series of research. The authors found that the same words can convey different emotional messages when accompanied by different tones or expressions. Driven by this motivation, they designed a network that improves the word representations by dynamically integrating the fine-grained representations of visual and auditory modalities into each word vector through a cross-modal gating mechanism, thereby achieving the goal of infusing non-verbal emotional information into word representations. For a word $\mathbf{L}^{(i)}$, the nonverbal shift vector $\mathbf{h}_m^{(i)}$ is calculated as follows:

$$\mathbf{h}_m^{(i)} = w_v^{(i)} \cdot (\mathbf{W}_v \mathbf{h}_v^{(i)}) + w_a^{(i)} \cdot (\mathbf{W}_a \mathbf{h}_a^{(i)}) + \mathbf{b}_h^{(i)}, \tag{20}$$

where $\mathbf{W}_v$ and $\mathbf{W}_a$ are weight matrices for the visual and acoustic embedding and $\mathbf{b}_h^{(i)}$ is the bias vector.

In [98], considering that audio and visual inputs might contain noise at certain time steps, like background noise in speech, the authors proposed a reinforcement learning-based gating unit to control the information fusion between fine-grained representations of different modalities. The gating mechanism allows for dynamic sentiment representation fusion by controlling whether the representation of the current word incorporates information from a particular modality. Unlike the previous two studies, which focus on capturing interactions of multimodal fine-grained sentiment representations associated with individual words, in [112], the authors modelled the feature interactions between multimodal fine-grained sentiment representations of multiple consecutive words within a window and used a memory neural network to model global information.

Pre-trained model-based fusion. Pre-trained language models have demonstrated strong language understanding capabilities, and researchers believe they also hold great potential for multimodal language understanding. To explore the capabilities of pre-trained language models in the field of multimodal sentiment analysis, in [99], the authors, inspired by the RAVEN method, designed a gating mechanism for pre-trained language models. The aim is to inject multimodal information into the intermediate layer word representations of the pre-trained language models to fully leverage their strong language modeling capabilities for efficient multimodal emotional understanding. In [113], the authors proposed a

cross-modal efficient attention mechanism that uses the output representations of pre-trained language models to compress the input sequences of visual and audio features, thereby enhancing the model's computational efficiency.

To effectively extend LLMs and LMMs to multimodal sentiment analysis tasks and address two major challenges in the field, namely, the low contribution rate of the visual modality and the design of an effective multimodal fusion architecture, scholars in [114] proposed the VLP2MSA model that integrates several novel components. Specifically, they introduced a fusion pipeline composed of a text encoder, a video encoder, a prompt module, a video-text contrastive learning module, and a multimodal integration encoder. Instead of using preprocessed visual features, their model directly takes raw text and video frames as input.

To better extract visual information, they developed the inter-frame hybrid transformer, which captures both facial expressions and body motion features from sparsely sampled video frames, thereby improving the representation of visual modality. In addition, a video-text prompting mechanism was designed to generate enhanced text representations by incorporating visual cues through cross-attention operations. This is intended to simulate the way humans intuitively combine spoken words with visual cues like facial expressions to detect emotions such as sarcasm. The final video-text representations are further aligned using a video-text contrastive learning strategy, which maps paired text and video embeddings into a shared space, ensuring semantic consistency before multimodal fusion. This fusion strategy aims to mitigate the adverse effects of modality heterogeneity and promote more accurate sentiment understanding.

3.3.3 *Audio-image-text sentiment analysis datasets*

We summarize the commonly used datasets for audio-image-text image-text sentiment analysis, including ICT-MMMO, IEMOCAP, CMU-MOSI, CMU-MOSEI, MELD, CH-SIMS, CH-SIMS 2, M3ED, MER2023, EMER, MER2024, CMU-MOSEAS and UR-FUNNY.

ICT-MMMO dataset is collected from the YouTube website and defines seven sentiment labels based on sentiment polarity and intensity: positive (strong), positive, positive (weak), neutral, negative (weak), negative, and negative (strong). In [88], the authors first addressed the task of tri-modal sentiment analysis and demonstrated that it is a feasible task that can benefit from the combined use of image, audio, and text modalities. This dataset forms the basis of their research.

IEMOCAP dataset is a multimodal video dialogue dataset collected by the SAIL lab at the University of Southern California. It contains about 12 h of multimodal data, including video, audio, facial motion capture, and transcribed text. The dataset was collected through dialogues by 5 professional male actors and 5 professional female actors in pairs, engaging in either improvised or scripted dialogues, with a focus on emotional expression. The dataset includes a total of 4787 improvised dialogues and 5255 scripted dialogues, with an average of 50 sentences per dialogue and an average duration of 4.5 seconds per sentence. Each sentence in the dialogue segments is annotated with specific emotional labels, divided into ten categories including anger, happiness, sadness, and neutral.

CMU-MOSI and **CMU-MOSEI** are two commonly used datasets in the multimodal sentiment analysis area. The data is sourced from video blogs (vlogs) on the online sharing platform YouTube. These datasets primarily focus on coarse-grained multimodal sentiment classification tasks. CMU-MOSI dataset comprises 2199 video segments extracted from 93 distinct videos. The video content consists of English comments posted by individual speakers. There are 41 female and 48 male speakers, mostly between the ages of 20 and 30, coming from diverse backgrounds (Caucasian, Asian, etc.). The videos are annotated by five annotators from the Amazon Mechanical Turk platform, and the annotations are averaged. Annotations cover seven categories of emotional tendencies ranging from -3 to $+3$. The CMU-MOSEI dataset is larger than the CMU-MOSI dataset, containing 23453 video segments from 1000 different speakers across 250 topics, with a total duration of 65 h. The dataset includes both emotion labels and sentiment labels. Emotion labels include happiness, sadness, anger, fear, disgust, and surprise, while sentiment labels include sentiment binary classification, five classification, and seven classification annotations.

MELD dataset originates from the classic TV series Friends. It comprises a total of 1443 dialogues and 13708 utterances, with an average of 9.5 sentences per dialogue and an average duration of 3.6 seconds per sentence. Each sentence in the dialogue segments is annotated with one of seven emotional labels, including anger, disgust, sadness, happiness, neutral, surprise, and fear. Additionally, each sentence is

also assigned a sentiment label, categorized as positive, negative, or neutral.

CH-SIMS dataset is a Chinese multimodal sentiment classification dataset with the unique feature of having both unimodal and multimodal sentiment labels. It consists of 60 original videos collected from movie clips, TV series, and various performance shows. These videos were clipped at the frame level to obtain 2281 video segments. Annotators labeled each video segment for four modalities: text, audio, silent video, and multimodal. To avoid cross-modal interference during annotation, annotators could only access information from the current modality. They first performed unimodal labeling, followed by multimodal labeling. Although the dataset provides labels for each modal, its primary purpose is coarse-grained multimodal sentiment classification. The **CH-SIMS 2** dataset expands the CH-SIMS dataset. This dataset is larger in scale and more difficult, requiring the model to accurately integrate information from different modalities to predict the correct answer.

M3ED dataset includes 990 Chinese dialogue videos, totaling 24449 sentences. Each sentence is annotated for six basic emotions (happiness, surprise, sadness, disgust, anger, and fear), as well as neutral emotion.

MER2023 includes four subsets: Train & Val, MERMULTi, MER-NOISE, and MER-SEMI. In the last subset, besides the labeled samples, it also contains a large amount of unlabeled data. The dataset annotates sentiment labels on each sample and focuses on challenges such as multi-label learning, noise robustness, and semi-supervised learning. Furthermore, they built upon MER2023 to create the **EMER** dataset, which not only annotates sentiment labels on each sample but also the reasoning process behind the labels. In **MER2024**, they expanded the dataset size and included a subset with multi-label annotations, attempting to describe the emotional states of characters as accurately as possible.

CMU-MOSEAS is the first large-scale multimodal language dataset for Spanish, Portuguese, German and French, with 40000 total labelled sentences. It covers a diverse set topics and speakers, and carries supervision of 20 labels including sentiment (and subjectivity), emotions, and attributes.

UR-FUNNY dataset is tailored for humor detection tasks, which are closely related to multimodal sentiment analysis. The dataset was collected from the TED website, selecting 8257 humorous snippets from 1866 videos and their transcribed texts, and additionally, 8257 non-humorous segments were randomly chosen. The total duration of the dataset is 90.23 h, encompassing 1741 different speakers and 417 distinct topics.

3.4 Multimodal sarcasm detection

Sarcasm detection task initially only focused on the textual context [145–147], with scholars noting that common ironic sentences often juxtapose positive phrases with negative contexts. For example, in the sentence "I'm so happy I'm late for work", the presence of the positive phrase "happy" within the negative context of being late for work makes it easily recognizable as sarcasm. In most cases, the sentiment signals conveyed by different modalities in multimodal data are consistent. However, there are instances of inconsistency, necessitating sentiment disambiguation across modalities. Multimodal sentiment disambiguation is essentially a classification task. Multimodal sentiment inconsistency can be categorized into two types: complete sentiment conflict, defined as multimodal irony recognition tasks, and instances where some modalities convey 'neutral' sentiment polarities while others convey positive or negative sentiment polarities, which are typical cases of implicit sentiment expression. The multimodal sarcasm detection task formalization as follows.

Multimodal sarcasm detection aims to identify if a given text associated with an image has sarcastic meaning. Formally, given a set of multimodal samples D, for each sample $d \in D$, it contains a sentence T with n words $\{t_1, t_2, t_3, \ldots, t_n\}$ and an associated image I. The goal of the model is to learn a multimodal sarcasm detection classifier to correctly predict the results of unseen samples.

In [120], the authors introduced a multimodal sarcasm detection task for videos and compiled a corresponding dataset from television series. Considering the correlation between sentiment classification and sarcasm detection, in [121], the authors proposed a multi-task framework to simultaneously recognize sarcasm and classify sentiment polarity. In [122], the authors suggested identifying sarcasm by capturing incongruent emotional semantic cues across modalities, such as rolling one's eyes while uttering praise.

Additionally, some researchers have studied sarcasm in text and images; for example, in [52], the authors introduced a multimodal sarcasm detection task for text and images and designed a multi-level fusion network to detect sarcasm. In [54], the authors proposed the multimodal sarcasm detection model using two different computational frameworks based on SVM and CNN that integrate text and visual modalities.

Identifying inconsistencies between modalities is key to multimodal sarcasm detection, and recent models can be categorized into those based on attention mechanisms and those using graph neural networks (GNN). For example, in [55], the authors introduced a BERT-based model with a cross-modal attention mechanism and a text-oriented co-attention mechanism to capture inconsistencies within and between modalities. In [56], the authors designed a 2D internal attention mechanism based on BERT and ResNet to extract relationships between words and images. In [57], the authors proposed a Transformer-based architecture to fuse textual and visual information. In terms of GNN, scholars in [58] built heterogeneous intramodal and cross-modal graphs (InCrossMG) for each multimodal example to determine the emotional inconsistencies within specific modalities and between different modalities, and introduced an interactive graph convolutional network structure to learn the relationships of inconsistencies in a joint and interactive manner within modal and cross-modal graphs. In [59], the authors constructed heterogeneous graphs containing fine-grained object information of images for each instance and designed a cross-modal graph convolutional network.

Additionally, some scholars [60] have proposed incorporating external commonsense knowledge into multimodal sarcasm detection to enhance the model's understanding of cross-modal semantics. They introduced a model named KnowleNet, which leverages the ConceptNet [148] knowledge base to enrich both textual and visual features with semantically related concepts. Unlike attention-only models, KnowleNet applies a knowledge-based word-level semantic similarity detection mechanism that compares text and image information based on their conceptual embeddings.

In this framework, the image is first parsed into a set of semantic attributes, while the accompanying text is tokenized. Both modalities are then mapped to a shared conceptual space via ConceptNet, producing word embeddings that capture commonsense relationships. For each word or attribute, ConceptNet generates dense vectors through a smoothed pointwise mutual information (PPMI) matrix followed by dimensionality reduction via SVD. To compute the semantic alignment between modalities, the model calculates inner product similarities between the text and image concept vectors, then applies a max pooling operation to extract the most salient associations, followed by a Flatten layer for feature unification. Crucially, the model adopts different processing procedures for sarcastic (positive) and non-sarcastic (negative) samples. For sarcastic samples, where image and text tend to conflict or misalign semantically, the model emphasizes semantic divergence; for non-sarcastic samples, it focuses on semantic consistency. To further refine representation, a contrastive learning strategy is employed to increase the separability of positive and negative samples in the shared representation space.

Recent studies have explored various strategies to enhance multimodal sarcasm detection by integrating external knowledge and exploiting richer modality interactions. For instance, the work in [61] proposed a lightweight multimodal interaction model that incorporates commonsense knowledge to improve deep learning-based sarcasm detection. Building upon this, Ref. [62] further introduced emotional knowledge into the detection process by leveraging sentiment dictionaries. Sentiment vectors are derived from words across different modalities and integrated into each modality's representation to capture affective signals more effectively. Advancing this line of research, Ref. [53] proposed a multi-view CLIP-based framework that extracts multi-grained features from textual, visual, and joint text-image interaction views, enabling a more comprehensive understanding of multimodal sarcasm. These studies reflect a clear progression, from incorporating general knowledge to emotional semantics, to fine-grained multimodal interactions, demonstrating the growing emphasis on leveraging diverse knowledge sources and multi-perspective cues for improved sarcasm understanding.

Building on the success of large-scale pretrained models, recent work has begun to explore the application of LLMs and LMMs for multimodal sarcasm understanding. In [63], the authors tested the performance of some existing open-source LLMs and LMMs in the multimodal sarcasm detection task and proposed a generative multi-media sarcasm model consisting of a designed instruction template and a demonstration retrieval module based on the large language model. In [64], the authors proposed a versatile framework named CofiPara for multimodal sarcasm target identification, following a coarse-to-fine paradigm. Each sample consists of an image-text pair $\{I, T\}$. The framework first performs coarse-grained sarcasm detection (MSD) to determine if a sample is sarcastic, then conducts fine-grained target identification (MSTI) to locate sarcasm targets in both text and image via textual phrases and visual bounding boxes. CofiPara integrates large multimodal model (LMM) reasoning to enhance performance, and includes three stages: divergent thinking with LMMs, coarse-grained pre-training, and fine-grained fine-tuning.

We summarize the commonly used datasets for multimodal sarcasm detection, including MMSD,

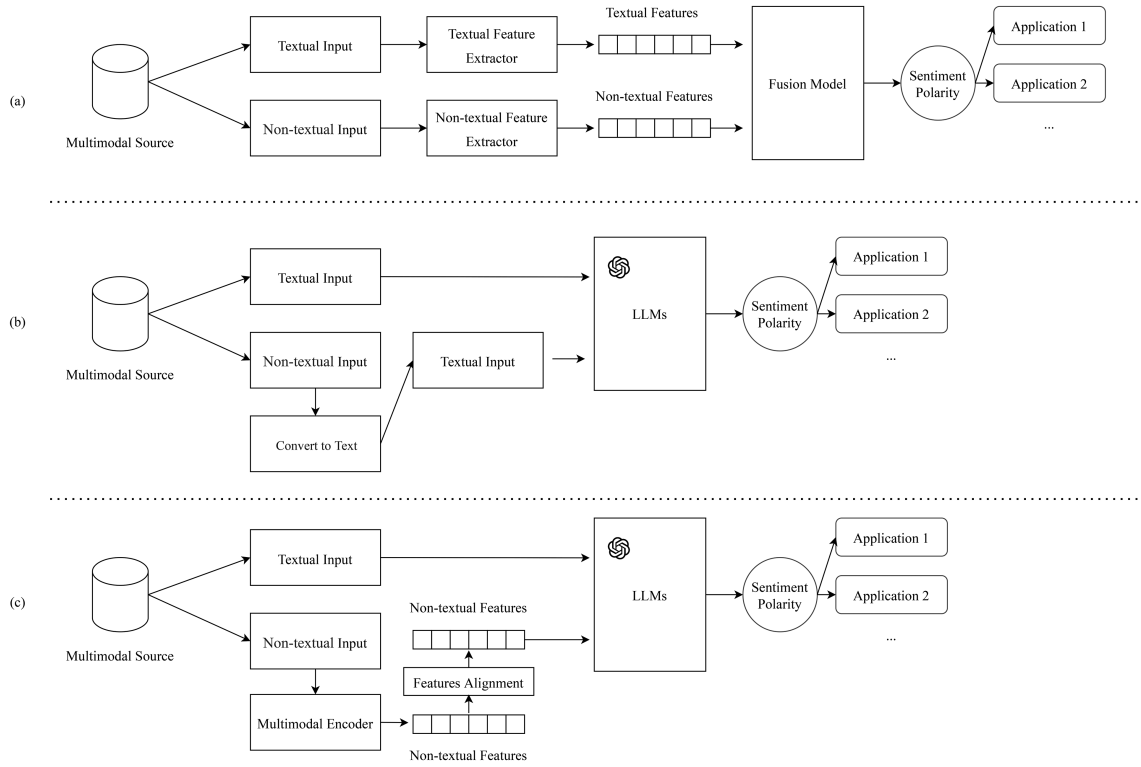

Figure 3 Conceptual illustration of multimodal sentiment analysis using LLMs. (a) LLMs are used as text feature extractors; (b) non-text modalities are converted into textual form via modality translation and then processed by LLMs; (c) non-text modality features are aligned to the textual representation space through modality alignment before being input into LLMs.

MMSD2.0 and MUStARD.

MMSD dataset is collected from the Twitter platform by searching for tweets in English that contain special tags indicating sarcasm, such as #sarcasm, #sarcastic, #irony, #ironic, to gather sarcastically labeled data, and collecting other tweets without these tags as non-sarcastic data. The dataset is annotated with a binary classification of "sarcastic/non-sarcastic". **MMSD2.0** fixed the shortcomings of MMSD, by removing the spurious cues and re-annotating the unreasonable samples.

MUStARD is a multimodal sarcasm detection dataset primarily sourced from English sitcoms, including *Friends*, *The Big Bang Theory*, *The Golden Girls*, and non-sarcastic video content from the MELD dataset. The authors collected a total of 6365 video clips from these sources and annotated them, including 345 sarcastic video clips. To balance the categories, an equal number of 345 non-sarcastic video clips were selected from the remaining clips, resulting in a dataset comprising 690 video segments. The annotations include the dialogue, speaker, context dialogue and its speaker, the source TV show, and a label indicating whether it is sarcastic. The rich annotation allows researchers to conduct a variety of learning tasks, including studying the impact of context and speakers on the task of sarcasm detection.

4 Usage of LLMs in multimodal sentiment analysis

Figure 3 illustrates three common strategies for leveraging LLMs in multimodal sentiment analysis tasks. These strategies aim to take advantage of the extensive knowledge and strong reasoning abilities of LLMs, while enabling them, despite being inherently text-oriented, to process and understand multimodal signals.

The first strategy, shown in Figure 3(a), resembles the traditional multimodal processing pipeline. Multimodal data is collected from sources such as social media platforms, then cleaned and filtered. Separate feature extraction algorithms are applied to each modality to generate corresponding feature vectors. These vectors are fused using multimodal fusion techniques and passed to classification models for sentiment prediction. In this setup, LLMs serve primarily as enhanced text feature extractors, without fundamentally altering the traditional architecture. The second strategy, depicted in Figure 3(b), is

Table 3 Some text-centric multimodal sentiment analysis methods that have utilized LLMs.

Method	Usage of LLMs	Advantage	Disadvantage
WisdoM [128]	Zero-shot learning: (1) using ChatGPT to provide prompt templates; (2) prompting LMMs to generate context using the prompt templates with image and sentence.	Leverage the contextual world knowledge induced from the LMMs for enhanced image-text sentiment classification.	Due to hallucinations in LLMs, the contextual knowledge supplemented by LLMs and LMMs as knowledge sources may not be accurate. The adaptive incorporation of context requires further exploration.
ChatGPT-ICL [87]	Zero-shot learning and few-shot learning: using ChatGPT to predict final sentiment labels.	This work explores the potential of ICL with ChatGPT for multimodal aspect-based sentiment analysis, achieves competitive performance while utilizing a significantly smaller sample size.	The ICL framework exhibits a relatively limited capability for aspect term extraction tasks when compared to fine-tuned methods.
A2II [77]	Full-parameter tuning: leverage the ability of LMMs to alleviate the limitation of cross-modal fusion.	This work explored an instruction tuning modeling approach for multimodal aspect-based sentiment classification task, and achieved impressive performance.	The visual features extracted by the Q-Former structure, which queries based on aspect, may be mismatched, leading to the neglect of some visual emotional signals.
CofiPara [64]	Zero-shot learning: using potential sarcastic labels as prompts to cultivate divergent thinking in LMMs, eliciting the relevant knowledge in LMMs for judging irony.	Note the negative impact of the inevitable noise in LMMs, and use competitive principles to align the sarcastic content generated by LMMs with their original multimodal features to reduce the noise impact. View LMMs as modal converters, transforming visual information into text to help cross-modal alignment.	Viewing LMMs as a knowledge source largely depends on the capabilities of the LMMs themselves. Although effective measures have been taken to reduce the impact of noise from LMMs, a certain proportion of erroneous judgments are still caused by LMMs.

modality translation, where non-text inputs are converted into textual descriptions, such as generating captions for images, so that the resulting textual representations can be directly processed by LLMs. The third strategy, illustrated in Figure 3(c), is modality alignment. Here, a multimodal encoder is used to extract features from various modalities, and a feature alignment mechanism maps these features into the LLM's textual representation space. This alignment allows the LLM to interpret and utilize multimodal information effectively for sentiment analysis.

In Table 3 [64, 77, 87, 128], we have summarized some representative multimodal sentiment analysis methods assisted by LLMs, analyzing the strategies used with LLMs as well as their advantages and disadvantages. After analysis, we have found that most existing research tends to view LLMs as knowledge sources. Operating under a parameter-fixed paradigm, these studies leverage zero-shot and few-shot strategies to endow smaller models with additional worldly knowledge in multimodal sentiment analysis tasks, resulting in performance improvements. Here are further advantages and methods of utilizing LLMs in text-centric multimodal sentiment analysis.

• LLMs can supplement richer knowledge, such as knowledge of different languages and cultures, to promote the progress of multimodal sentiment analysis towards multilingualism.

• Leveraging the robust multimodal capabilities of LMMs, models like GPT-4V and LLaVA, known for their strong image captioning abilities, can transform image data into textual format, simplifying the challenge of modal alignment.

• Utilizing the powerful reasoning capabilities of LLMs, existing work has shown that effective in-context learning (ICL) can enhance the emotional reasoning capabilities of LLMs, significantly improving their ability to trace and guide emotional understanding.

• Fine-tuning with high-quality multimodal sentiment data using a parameter-tuning paradigm, such as the A^2II model, has also been successful. Although it used the smaller-scale Flan-T5-base model, there is anticipation for methods that adopt parameter-efficient fine-tuning strategy in larger-scale LLMs.

• Additionally, the use of LLMs as tools in multimodal sentiment analysis holds a promising outlook.

However, there are also disadvantages of using LLMs in multimodal sentiment analysis.

• LLMs have to face hallucination problems, and the inevitable generation of erroneous knowledge may lead to incorrect judgments. Enhancing the accuracy and completeness of sentiment judgment-related

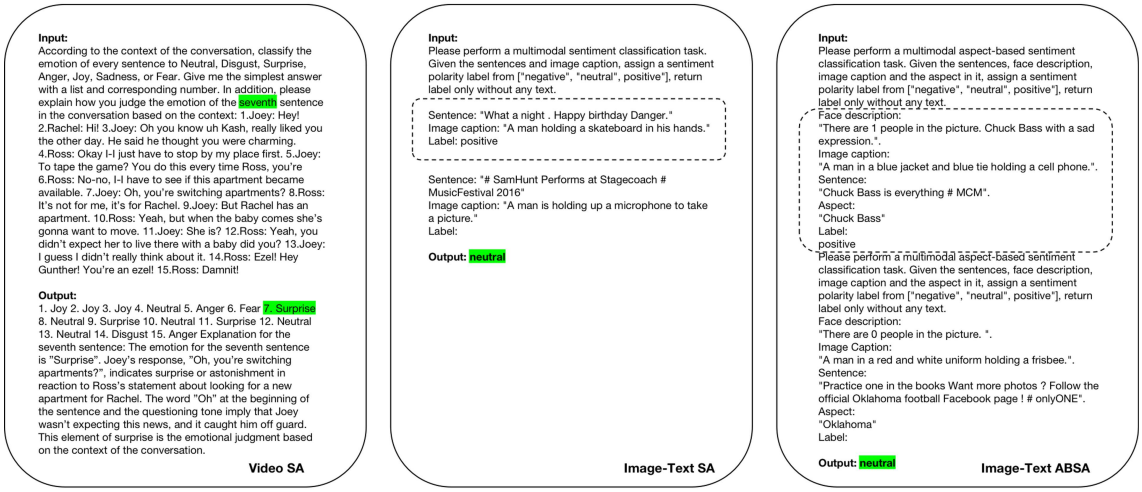

Figure 4 (Color online) Prompt examples for video-based sentiment analysis (video SA), image-text sentiment classification (image-text SA), and multimodal aspect-based sentiment classification (image-text ABSA), respectively. The text inside the dashed box is a demonstration of the few-shot setting and would be removed under the zero-shot setting.

knowledge from LLMs while reducing the negative noise caused by biases and hallucinations remains a pressing challenge.

• The sensitivity of LLMs to prompts is significant, as different prompts can drastically influence the output. Choosing the appropriate prompt is challenging.

• Not all LLMs excel in emotional intelligence; as the training of LLMs and LMMs currently aims to develop a broad range of capabilities, emotional intelligence is just one of many focal points. Therefore, the emotional capabilities of most models may not be exceptional, and careful consideration is needed when selecting LLMs for assisting in multimodal sentiment analysis.

• Existing LMMs still lack support for additional modalities. While most LMMs focus on text and image modalities, and some have video processing capabilities, there is a lack of capacity to handle other modalities like physiological signals, limiting their use in multimodal sentiment analysis.

• Methods based on the parameter-tuning paradigm face significant costs, requiring several times the computational resources and time compared to traditional multimodal sentiment analysis models.

5 Evaluations of LLMs-based multimodal sentiment analysis methods

5.1 Prompting strategy

When using LLMs, we employ prompts (a specific type of input text) to trigger the model's response. Since LLMs are highly sensitive to prompts, even slight variations in semantics can elicit vastly different responses. Therefore, prompt design is of paramount importance. Figure 4 shows some prompt examples.

As shown in Figure 4, in the zero-shot setting, the prompts include the task name, task definition, and output format. The task name is used to identify and specify the task, while the task definition provides an explanation of the task, enabling the model to understand the input-output format of the task and providing a candidate label space for outputs. The output format defines the expected structure of the output, guiding the model to generate content in the expected format.

In the few-shot setting, additional demonstration sections are added to assist in model inference learning.

5.2 Evaluation metrics

This section discusses the various commonly used metrics used in the field of text-centric multimodal sentiment analysis tasks [149].

Accuracy is a measure that indicates the proportion of instances correctly predicted out of the total number of instances. Further, **Weighted-Accuracy** accounts for class imbalances by assigning different

weights to each class.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{TN}}, \tag{21}$$

$$\text{Weighted-Accuracy} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{TN}_i}, \tag{22}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, FN represents false negatives, N represents the total number of instances, w_i represents the weight for class i.

Precision evaluates the fraction of true positive predictions among instances that have been predicted as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{23}$$

Recall, which is sometimes referred to as sensitivity or true positive rate, quantifies the proportion of true positive instances that are accurately predicted.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{24}$$

The **F1-Score** merges precision and recall to offer a well-rounded assessment of the model's accuracy. Additionally, the **Weighted-F1-Score** accounts for class imbalances.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{25}$$

$$\text{Weighted-F1-Score} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}. \tag{26}$$

5.3 Reference results

With the in-depth development of LLMs in the field of multimodal sentiment analysis, it is necessary to compare the performance of LLMs on multimodal sentiment analysis datasets. However, testing on commercial LLMs such as ChatGPT is often expensive. Some studies [63, 64, 77, 87, 128, 150–152] have demonstrated the performance of some LLMs on multimodal sentiment analysis tasks. We have organized the relevant results in the Table 4.

6 Applications of text-centric multimodal sentiment analysis

The research in text-centric multimodal sentiment analysis has its roots in the flourishing development of multimodal data and the advancements in deep learning technologies. It is also driven by a wide range of practical applications. In this section, we explore the application of LLM-based text-centric multimodal sentiment analysis.

6.1 Comment analysis

One of the earliest and most impactful applications of sentiment analysis was in the field of e-commerce for comment analysis. This research area not only attracted numerous computer scientists who delved into algorithm development but also drew the interest of management scientists exploring marketing and management strategies. Initially, these studies primarily revolved around textual comments, analyzing user reviews to gather feedback on products or services. However, as e-commerce evolved, relying solely on text-based sentiment analysis proved insufficient. User-generated comments often include multimedia elements, making multimodal data more prominent compared to pure text comments.

With the increasing availability of multimodal data on social networks, some of the challenges that puzzled researchers can be alleviated in a multimodal interactive context, enabling comprehensive sentiment analysis. For instance, one challenging problem is sarcasm recognition, which can be easily resolved

Table 4 An overview of the performance (%) of existing LLMs and LMMs on text-centric multimodal sentiment analysis benchmarks. $*$ indicates the model results after training on the MMSD and MMSD2.0 datasets. Italicized words represent the few-shot results, and the rest are zero-shot results. Bold represents the best zero-shot results.

Method	MVSA-S	MVSA-M	TumEmo	Twitter-2015		Twitter-2017		MMSD		MMSD2.0		MOSI-2		MOSEI		CH-SIMS		M^{3}ED	
	Acc.	Acc.	Acc.	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
ChatGPT	56.55	53.18	48.17	65.28	**63.12**	62.47	**62.34**	69.02	–	–	–	86.13	85.92	85.60	84.43	79.66	78.78	44.47	**40.40**
Claude	–	–	–	–	–	–	–	–	–	–	–	87.04	**86.55**	85.83	**84.81**	**88.70**	**87.44**	34.90	34.83
LLaMA1-7B	67.23	60.72	38.26	58.53	–	46.43	–	58.99	–	–	–	82.01	–	75.62	–	–	–	–	–
LLaMA1-13B	66.88	68.82	44.68	52.07	–	47.24	–	57.53	–	–	–	72.10	–	79.55	–	–	–	–	–
LLaMA2-7B	66.99	69.22	40.28	58.53	–	46.60	–	56.33	–	–	–	67.68	–	77.30	–	–	–	–	–
LLaMA2-13B	66.02	68.69	45.78	60.37	–	48.54	–	60.23	–	–	–	81.86	–	81.66	–	–	–	–	–
Mixtral-AWQ	54.37	55.59	43.94	55.45	–	60.21	–	64.38	–	–	–	87.92	–	79.30	–	–	–	–	–
Gemma	67.72	61.61	43.10	54.29	–	52.43	–	60.07	–	–	–	81.65	–	77.05	–	–	–	–	–
Flan-T5-XXL	64.81	66.01	49.56	**72.13**	–	63.70	–	71.40	–	–	–	89.60	–	86.52	–	–	–	–	–
ChatGLM2-6B	–	–	–	–	–	–	–	–	–	–	–	84.12	84.12	–	–	77.58	75.95	**45.68**	30.52
ChatGLM2-6B*	–	–	–	–	–	–	–	94.02	93.76	78.41	78.23	–	–	–	–	–	–	–	–
LLaMA2-7B*	–	–	–	–	–	–	–	93.97	93.72	82.52	82.27	–	–	–	–	–	–	–	–
GPT-4V	76.19	**71.05**	50.58	53.85	–	60.16	–	**76.76**	–	–	–	**90.91**	–	87.10	–	–	–	–	–
Claude3-V	**80.95**	69.08	46.15	38.46	–	54.47	–	71.37	–	–	–	78.79	–	79.93	–	–	–	–	–
Gemini-V	72.73	70.18	51.65	54.51	–	59.32	–	56.83	–	–	–	88.34	–	**87.14**	–	–	–	–	–
OpenFlamingo	55.58	61.15	29.47	57.28	–	46.19	–	52.68	–	–	–	79.97	–	77.30	–	–	–	–	–
Fromage	29.85	28.19	22.76	19.96	–	27.31	–	40.68	–	–	–	57.19	–	47.41	–	–	–	–	–
LLaVA-v0-7B	69.42	65.42	30.44	35.10	–	44.57	–	43.21	–	–	–	74.69	–	74.65	–	–	–	–	–
LLaVA-v0-13B	73.06	69.61	38.51	37.99	–	48.46	–	44.29	–	–	–	80.18	–	76.58	–	–	–	–	–
LLaVA-v1.6-7B	59.95	67.23	45.12	59.31	–	52.84	–	59.61	–	–	–	85.63	–	81.62	–	–	–	–	–
LLaVA-v1.6-13B	64.56	60.43	53.22	58.73	–	56.08	–	62.31	–	–	–	86.39	–	78.26	–	–	–	–	–
MiniGPT4	71.12	70.78	50.29	47.16	–	49.43	–	57.49	–	–	–	83.99	–	83.38	–	–	–	–	–
mPLOG-Owl	51.94	50.36	33.37	33.75	–	38.74	–	49.73	–	–	–	68.75	–	58.10	–	–	–	–	–
mPLOG-Owl2.1	53.64	63.11	47.02	60.66	–	55.11	–	60.48	–	–	–	85.63	–	73.47	–	–	–	–	–
AdapterV2	73.54	70.13	39.14	37.32	–	48.38	–	57.20	–	–	–	86.43	–	82.02	–	–	–	–	–
VPGTrans	64.32	69.54	46.17	42.62	–	44.81	–	65.04	–	–	–	76.22	–	76.76	–	–	–	–	–
MultiGPT	52.91	62.03	30.26	58.53	–	46.35	–	59.82	–	–	–	68.35	–	72.76	–	–	–	–	–
LaVIN-7B	39.32	40.75	26.84	37.22	–	33.06	–	60.48	–	–	–	71.41	–	69.97	–	–	–	–	–
LaVIN-13B	53.64	48.79	32.77	35.39	–	40.68	–	57.58	–	–	–	79.97	–	73.54	–	–	–	–	–
Lynx	64.32	67.71	42.79	46.00	–	47.00	–	43.96	–	–	–	74.77	–	73.72	–	–	–	–	–
Fuyu-8B	48.54	55.46	46.34	58.82	–	50.81	–	61.44	–	–	–	83.49	–	78.37	–	–	–	–	–
LaVIT	61.65	68.74	41.78	36.84	–	43.36	–	56.00	–	–	–	73.09	–	64.10	–	–	–	–	–
Qwen-VL-Chat	62.38	69.06	49.29	65.48	–	59.72	–	61.10	–	–	–	85.93	–	80.41	–	–	–	–	–
BLIP	66.26	68.22	51.06	70.78	–	**64.42**	–	72.02	–	–	–	88.99	–	86.88	–	–	–	–	–
InstructBLIP	71.60	70.37	**52.36**	57.57	59.63	60.37	35.96	73.10	–	–	–	88.68	–	85.98	–	–	–	–	–
LLaVA-v1.5-7B*	–	–	–	–	–	–	–	93.67	93.40	85.18	85.11	–	–	–	–	–	–	–	–
mPLUG-Owl2	–	–	–	*76.80*	*72.30*	*74.20*	*73.00*	–	–	–	–	–	–	–	–	–	–	–	–
MMICL-14B	–	–	–	*76.00*	*72.70*	*74.10*	*74.00*	–	–	–	–	–	–	–	–	–	–	–	–
LLaVA-v1.5-13B	–	–	–	*77.90*	*74.30*	*74.60*	*74.30*	–	–	51.06	43.02	–	–	–	–	–	–	–	–
Qwen-VL-v1.0	–	–	–	–	–	–	–	–	–	**76.63**	**69.03**	–	–	–	–	–	–	–	–

with the addition of multimodal information. For example, when a comment like "It's such a surprise" is accompanied by a picture of a disappointed face, sarcasm recognition becomes straightforward. In the field of management, multimodal data, enriched with additional modal factors, can influence user decisions and consequently impact marketing and management strategies.

In practical applications, fine-grained sentiment analysis is more effective. In text-based analysis studies, user textual comment data can be broken down into fine-grained segments (e.g., sentences, clauses), with each segment evaluating different aspects of the main entity (e.g., price, quality, appearance). In contrast, fine-grained analysis of multimodal data is still in its emerging stages but presents greater challenges. For instance, extracting object-level information from complex multimodal data and modeling fine-grained element correspondences between multimodal elements are ongoing research topics that need exploration with LLMs in the future.

6.2 Multimodal intelligent human-machine interaction

Multimodal sentiment analysis can be applied to human-computer interaction, enabling real-time understanding and analysis of emotional communication for more natural interactions. There are three main categories of applications.

(1) Customer service conversations. In this domain, multimodal data consists of audio data and text data transformed from automatic speech recognition (ASR) technology. It mainly serves two tasks: customer satisfaction analysis and detection of customer abnormal emotions. Customer satisfaction analysis involves using multimodal emotion computing technology to analyze the content of conversations between customers and service representatives to assess the level of customer satisfaction. Customer abnormal emotion detection monitors customer emotions in real time through the analysis of customer dialogue data and prompts timely intervention when abnormal emotional changes occur.

(2) Emotional companionship. Emotional companionship is a crucial aspect of chatbot applications. Currently, most companion chatbots do not utilize multimodal emotion computing technology, meaning they do not fully possess human-like multimodal processing capabilities. Ideally, companion chatbots should be capable of recognizing and generating multimodal emotional features, such as expressing emotions through language, exhibiting emotional fluctuations in speech, or displaying facial expressions.

(3) Smart furniture. The development of artificial intelligence has given rise to smart homes, which enhance convenience and comfort in daily life and are increasingly popular among consumers. Many tech companies worldwide have entered the smart home market, proposing a range of solutions like Apple's HomeKit, Xiaomi's Mi Home, and Haier's U-Home. While smart homes have made life more convenient, they are currently primarily focused on home automation, with users controlling home devices through voice commands based on keyword recognition. This approach does not fully embody the intelligence of smart homes, and there is substantial potential for further development, particularly in voice interaction and automatic environment detection.

With the assistance of LLMs, AI technologies based on multimodal sentiment analysis methods can elevate the intelligence of smart homes in the future. True smart home scenarios involve multiple modalities, where smart home products can provide appropriate feedback by calculating the user's emotions (e.g., happiness, anger, sadness) or states (e.g., fatigue, restlessness). For example, based on a user's fatigue state in a multimodal scenario, the system could ask if they want the lights dimmed. In conversational scenarios, the system can detect the user's emotional state and provide empathetic responses. Smart in-car systems can promptly detect abnormal user emotions or states (e.g., road rage, fatigue) and provide appropriate reminders. Designing these functionalities poses significant challenges, but it also represents a significant opportunity for multimodal emotion computing to enter the smart furniture domain.

7 Conclusion

In this survey, we introduced the latest advancements in text-centric multimodal sentiment analysis area and summarized the primary challenges and potential solutions. Additionally, we reviewed the existing ways of applying LLMs in multimodal sentiment analysis tasks and summarized their advantages and disadvantages. We believe that leveraging LLMs in multimodal sentiment analysis has several potential advantages. (1) Knowledge source: LLMs trained on massive datasets and can be treated as a knowledge source, that can capture a broader range of patterns, linguistic cues, and contextual information related to emotions, potentially improving recognition performance. (2) Interpretability: LLMs can potentially elucidate the reasoning behind their decisions, enhancing the interpretability and transparency of the emotion recognition process. (3) Cross-domain applications: LLMs have the potential to be applied across various domains, as they are trained on a wide range of data sources. This allows them to understand emotions expressed in various domains, from customer reviews to conversational data, thus achieving broader applicability. However, LLMs-based methods also have to face problems such as hallucinations and high fine-tuning costs. Furthermore, as new models and datasets continue to emerge, ensuring the relevance and reliability of multimodal sentiment analysis requires continuous benchmarking and dataset expansion. Establishing standardized evaluation frameworks will be essential for tracking progress and maintaining robustness in this rapidly evolving field. The emergence of LLMs provides new ideas and challenges for multimodal sentiment analysis. We hope that this survey can help and encourage further research in this field.

Acknowledgements This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (Grant No. 2023ZD0121100) and National Natural Science Foundation of China (Grant Nos. 62441614, 62176078). We thank the anonymous reviewers for their insightful comments and suggestions.

References

1 Zhang C, Yang Z, He X, et al. Multimodal intelligence: representation learning, information fusion, and applications. IEEE J Sel Top Signal Process, 2020, 14: 478–493

2 Soleymani M, Garcia D, Jou B, et al. A survey of multimodal sentiment analysis. Image Vision Comput, 2017, 65: 3–14

3 Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2020

4 Rae J W, Borgeaud S, Cai T, et al. Scaling language models: methods, analysis & insights from training Gopher. 2021. ArXiv:2112.11446

5 OpenAI. GPT-4 technical report. 2023. ArXiv:2303.08774

6 Zhong Q, Ding L, Liu J, et al. Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. 2023. ArXiv:2302.10198

7 Bang Y, Cahyawijaya S, Lee N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. 2023. ArXiv:2302.04023

8 Ye J, Chen X, Xu N, et al. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. 2023. ArXiv:2303.10420

9 Yang J, Jin H, Tang R, et al. Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. 2023. ArXiv:2304.13712

10 Deng X, Bashlovkina V, Han F, et al. LLMs to the moon? Reddit market sentiment analysis with LLMs. In: Proceedings of the ACM Web Conference (WWW), 2023. 1014–1019

11 Wang Z, Xie Q, Ding Z, et al. Is ChatGPT a good sentiment analyzer? A preliminary study. 2023. ArXiv:2304.04339

12 Zhang W, Deng Y, Liu B, et al. Sentiment analysis in the era of large language models: a reality check. 2023. ArXiv:2305.15005

13 Zhao W, Zhao Y, Lu X, et al. Is ChatGPT equipped with emotional dialogue capabilities? 2023. ArXiv:2304.09582

14 Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. 2022. ArXiv:2204.02311

15 Taylor R, Kardas M, Cucurull G, et al. Galactica: a large language model for science. 2022. ArXiv:2211.09085

16 Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. 2023. arXiv:2302.13971

17 Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners. In: Proceedings of the 10th International Conference on Learning Representations (ICLR), 2022

18 Sanh V, Webson A, Raffel C, et al. Multitask-prompted training enables zero-shot task generalization. In: Proceedings of the 10th International Conference on Learning Representations (ICLR), 2022

19 Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. 2022. ArXiv:2210.11416

20 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2017

21 Zhang Z, Peng L, Pang T, et al. Refashioning emotion recognition modelling: the advent of generalised large models. 2023. ArXiv:2308.11578

22 Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models. 2023. ArXiv:2312.11805

23 Girdhar R, El-Nouby A, Liu Z, et al. ImageBind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 15180–15190

24 Li J, Li D, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the International Conference on Machine Learning (ICML), 2023. 19730–19742

25 Liu H, Li C, Wu Q, et al. Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2024

26 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML), 2021. 8748–8763

27 Chiang W L, Li Z, Lin Z, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90% ChatGPT quality. 2023. https://vicuna.lmsys.org

28 Sharma P, Ding N, Goodman S, et al. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018. 2556–2565

29 Lu P, Mishra S, Xia T, et al. Learn to explain: multimodal reasoning via thought chains for science question answering. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2022. 2507–2521

30 Bai J, Bai S, Yang S, et al. Qwen-VL: a frontier large vision-language model with versatile abilities. 2023. ArXiv:2308.12966

31 Qin L, Chen Q, Feng X, et al. Large language models meet NLP: a survey. 2024. ArXiv:2405.12819

32 Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of the International Conference on Machine Learning (ICML), 2019. 2790–2799

33 Hu E J, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. 2021. ArXiv:2106.09685

34 Li X L, Liang P, et al. Prefix-tuning: optimizing continuous prompts for generation. 2021. ArXiv:2101.00190

35 Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: efficient finetuning of quantized LLMs. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2024

36 Niu T, Zhu S, Pang L, et al. Sentiment analysis on multi-view social data. In: Proceedings of the 22nd International Conference on MultiMedia Modeling, 2016. 15–27

37 Ramamoorthy S, Gunti N, Mishra S, et al. Memotion 2: dataset on sentiment and emotion analysis of memes. In: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, 2022

38 Jia A, He Y, Zhang Y, et al. Beyond emotion: a multi-modal dataset for human desire understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2022. 1512–1522

39 Truong Q T, Lauw H W. VistaNet: visual aspect attention network for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 305–312

40 Borth D, Ji R, Chen T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia, 2013. 223–232

41 Wang M, Cao D, Li L, et al. Microblog sentiment analysis based on cross-media bag-of-words model. In: Proceedings of the International Conference on Internet Multimedia Computing and Service, 2014

42 Cai G, Xia B. Convolutional neural networks for multimedia sentiment analysis. In: Proceedings of National CCF Conference on Natural Language Processing and Chinese Computing (NLPCC), 2015. 159–167

43 Yu Y, Lin H, Meng J, et al. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms, 2016, 9: 41

44 Xu N. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI), 2017. 152–154

45 Xu N, Mao W. MultiSentiNet: a deep semantic network for multimodal sentiment analysis. In: Proceedings of the ACM Conference on Information and Knowledge Management (CIKM), 2017. 2399–2402

46 Xu N, Mao W, Chen G, et al. A co-memory network for multimodal sentiment analysis. In: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2018. 929–932

47 Yang X, Feng S, Zhang Y, et al. Multimodal sentiment detection based on multi-channel graph neural networks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. 328–339

48 Yu Y, Zhang D, Li S, et al. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In: Proceedings of the 30th ACM International Conference on Multimedia, 2022. 189–198

49 Yu Y, Zhang D. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2022. 1–6

50 Zhu T, Li L, Yang J, et al. Multimodal emotion classification with multi-level semantic reasoning network. IEEE Trans Multimedia, 2023, 25: 6868–6880

51 Yang X, Feng S, Wang D, et al. Image-text multimodal emotion classification via multi-view attentional network. IEEE Trans Multimedia, 2020, 23: 4014–4026

52 Cai Y, Cai H, Wan X, et al. Multimodal sarcasm detection in Twitter with hierarchical fusion model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019. 2506–2515

53 Qin L, Huang S, Chen Q, et al. MMSD 2.0: towards a reliable multi-modal sarcasm detection system. 2023. ArXiv:2307.07135

54 Schifanella R, de Juan P, Tetreault J, et al. Detecting sarcasm in multimodal social platforms. In: Proceedings of the 24th ACM International Conference on Multimedia, 2016. 1136–1145

55 Pan H, Lin Z, Fu P, et al. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In: Proceedings of Findings of the Association for Computational Linguistics, 2020. 1383–1392

56 Wang X, Sun X, Yang T, et al. Building a bridge: a method for image-text sarcasm detection without pretraining on image-text data. In: Proceedings of the 1st International Workshop on Natural Language Processing Beyond Text, 2020. 19–29

57 Tomás D, Ortega-Bueno R, Zhang G, et al. Transformer-based models for multimodal irony detection. J Ambient Intell Hum Comput, 2023, 14: 7399–7410

58 Liang B, Lou C, Li X, et al. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021. 4707–4715

59 Liang B, Lou C, Li X, et al. Multi-modal sarcasm detection via cross-modal graph convolutional network. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022. 1767–1777

60 Yue T, Mao R, Wang H, et al. KnowleNet: knowledge fusion network for multimodal sarcasm detection. Inf Fusion, 2023, 100: 101921

61 Liu H, Yang B, Yu Z. A multi-view interactive approach for multimodal sarcasm detection in social Internet of Things with knowledge enhancement. Appl Sci, 2024, 14: 2146

62 Fu H, Liu H, Wang H, et al. Multi-modal sarcasm detection with sentiment word embedding. Electronics, 2024, 13: 855

63 Tang B, Lin B, Yan H, et al. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2024. 1732–1742

64 Lin H, Chen Z, Luo Z, et al. CofiPara: a coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. 2024. ArXiv:2405.00390

65 Yu J, Jiang J. Adapting BERT for target-oriented multimodal sentiment classification. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2019

66 Zheng Z, Zhang Z, Wang Z, et al. Decompose, prioritize, and eliminate: dynamically integrating diverse representations for multimodal named entity recognition. In: Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 2024. 4498–4508

67 Wu Z, Zheng C, Cai Y, et al. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020. 1038–1046

68 Zhang Q, Fu J, Liu X, et al. Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2018

69 Sun L, Wang J, Zhang K, et al. RpBERT: a text-image relation propagation-based BERT model for multimodal NER. In:

Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021. 13860–13868
70 Yu J, Jiang J, Yang L, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: Proceedings of the Association for Computational Linguistics (ACL), 2022
71 Moon S, Neves L, Carvalho V, et al. Multimodal named entity recognition for short social media posts. 2018. ArXiv:1802.07862
72 Zhang D, Wei S, Li S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021. 14347–14355
73 Xu N, Mao W, Chen G, et al. Multi-interactive memory network for aspect-based multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2019. 371–378
74 Khan Z, Fu Y, et al. Exploiting BERT for multimodal target sentiment classification through input space translation. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021. 3034–3042
75 Yu J, Jiang J, Xia R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE ACM Trans Audio Speech Lang Process, 2020, 28: 429–439
76 Yang H, Zhao Y, Qin B, et al. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022. 3324–3335
77 Feng J, Lin M, Shang L, et al. Autonomous aspect-image instruction (A2II): Q-Former guided multimodal sentiment classification. In: Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 2024. 1996–2005
78 Ling Y, Yu J, Xia R, et al. Vision-language pre-training for multimodal aspect-based sentiment analysis. 2022. ArXiv:2204.07955
79 Ju X, Zhang D, Xiao R, et al. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021. 4395–4405
80 Yang L, Na J C, Yu J. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. Inf Process Manage, 2022, 59: 103038
81 Xiao L, Wu X, Xu J, et al. Atlantis: aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. Inf Fusion, 2024, 106: 102304
82 Zhao F, Li C, Wu Z, et al. M2DF: multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis. 2023. ArXiv:2310.14605
83 Zhou R, Guo W, Liu X, et al. AoM: detecting aspect-oriented information for multimodal aspect-based sentiment analysis. 2023. ArXiv:2306.01004
84 Liu Y, Zhou Y, Li Z, et al. RNG: reducing multi-level noise and multi-grained semantic GAP for joint multimodal aspect-sentiment analysis. 2024. ArXiv:2405.13059
85 Li Y, Ding H, Lin Y, et al. Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. Artif Intell Rev, 2024, 57: 1–26
86 Yang X, Feng S, Wang D, et al. Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. 2023. ArXiv:2305.10169
87 Yang L, Wang Z, Li Z, et al. An empirical study of multimodal entity-based sentiment analysis with ChatGPT: improving in-context learning via entity-aware contrastive learning. Inf Process Manage, 2024, 61: 103724
88 Morency L P, Mihalcea R, Doshi P, et al. Towards multimodal sentiment analysis: harvesting opinions from the web. In: Proceedings of the 13th International Conference on Multimodal Interfaces, 2011. 169–176
89 Zadeh A, Zellers R, Pincus E, et al. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. 2016. ArXiv:1606.06259
90 Zadeh A B, Liang P P, Poria S, et al. 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. 2236–2246
91 Zadeh A, Cao Y S, Hessner S, et al. CMU-MOSEAS: a multimodal language dataset for Spanish, Portuguese, German and French. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020
92 Yu W, Xu H, Meng F, et al. CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotations of modality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020. 3718–3727
93 Liu Y, Yuan Z, Mao H, et al. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), 2022. 247–258
94 Poria S, Hazarika D, Majumder N, et al. MELD: a multimodal multi-party dataset for emotion recognition in conversations. 2018. ArXiv:1810.02508
95 Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017. 1103–1114
96 Wang Y, Shen Y, Liu Z, et al. Words can shift: dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 7216–7223
97 Hazarika D, Zimmermann R, Poria S, et al. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020. 1122–1131

98 Chen M, Wang S, Liang P P, et al. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017. 163–171

99 Rahman W, Hasan M K, Lee S, et al. Integrating multimodal information in large pretrained transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020. 2359–2369

100 Li L, Chen Y C, Cheng Y, et al. HERO: hierarchical encoder for video+ language omni-representation pre-training. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 2046–2065

101 Cao D, Ji R, Lin D, et al. A cross-media public sentiment analysis system for microblog. Multimedia Syst, 2016, 22: 479–486

102 Yu W, Xu H, Yuan Z, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. 2021. ArXiv:2102.04830

103 Wu Y, Zhao Y, Yang H, et al. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors. 2022. ArXiv:2203.00257

104 Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019. 6558–6569

105 Huang J, Pu Y, Zhou D, et al. Multimodal sentiment analysis based on 3D stereoscopic attention. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. 11151–11155

106 Mai S, Zeng Y, Zheng S, et al. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Trans Affective Comput, 2023, 14: 2276–2289

107 Lin R, Hu H. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. 2022. ArXiv:2210.14556

108 Sun Z, Sarma P, Sethares W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020. 8992–8999

109 Mai S, Hu H, Xing S, et al. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020. 164–172

110 Liu Z, Shen Y, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors. 2018. ArXiv:1806.00064

111 Mai S, Hu H, Xing S, et al. 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 (ACL), 2019. 481–492

112 Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2018

113 Dai W, Cahyawijaya S, Liu Z, et al. Multimodal end-to-end sparse model for emotion recognition. 2021. ArXiv:2103.09666

114 Yi G, Fan C, Zhu K, et al. VLP2MSA: expanding vision-language pre-training to multimodal sentiment analysis. Knowl-Based Syst, 2024, 283: 111136

115 Busso C, Bulut M, Lee C C, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Evaluation, 2008, 42: 335–359

116 Zhao J, Zhang T, Hu J, et al. M3ED: multi-modal multi-scene multi-label emotional dialogue database. 2022. ArXiv:2205.10237

117 Lian Z, Sun H, Sun L, et al. MER 2023: multi-label learning, modality robustness, and semi-supervised learning. In: Proceedings of the 31st ACM International Conference on Multimedia, 2023. 9610–9614

118 Lian Z, Sun L, Xu M, et al. Explainable multimodal emotion reasoning. 2023. ArXiv:2306.15401

119 Lian Z, Sun H, Sun L, et al. MER 2024: semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. 2024. ArXiv:2404.17113

120 Castro S, Hazarika D, Pérez-Rosas V, et al. Towards multimodal sarcasm detection (an obviously perfect paper). In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019. 4619–4629

121 Chauhan D S, Dhanush S R, Ekbal A, et al. Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020. 4351–4360

122 Wu Y, Zhao Y, Lu X, et al. Modeling incongruity between modalities for multimodal sarcasm detection. IEEE MultiMedia, 2021, 28: 86–95

123 Chen L, Zhang H, Xiao J, et al. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 6298–6306

124 Zhao H, Yang M, Bai X, et al. A survey on multimodal aspect-based sentiment analysis. IEEE Access, 2024, 12: 12039–12052

125 Hasan M K, Rahman W, Zadeh A, et al. UR-FUNNY: a multimodal language dataset for understanding humor. 2019. ArXiv:1904.06618

126 Thelwall M, Buckley K, Paltoglou G, et al. Sentiment strength detection in short informal text. J Am Soc Inf Sci, 2010, 61: 2544–2558

127 Li Z, Xu B, Zhu C, et al. CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. 2022. ArXiv:2204.05515

128 Wang W, Ding L, Shen L, et al. WisdoM: improving multimodal sentiment analysis by fusing contextual world knowledge. 2024. ArXiv:2401.06659

129 Ma D, Li S, Wu F, et al. Exploring sequence-to-sequence learning in aspect term extraction. In: Proceedings of the 57th

Annual Meeting of the Association for Computational Linguistics (ACL), 2019. 3538–3547
130 Chen Z, Qian T. Enhancing aspect term extraction with soft prototypes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 2107–2117
131 Karamanolakis G, Hsu D, Gravano L, et al. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. 2019. ArXiv:1909.00415
132 Peng T, Li Z, Zhang L, et al. FSUIE: a novel fuzzy span mechanism for universal information extraction. 2023. ArXiv:2306.14913
133 Peng T, Li Z, Wang P, et al. A novel energy based model mechanism for multi-modal aspect-based sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024. 18869–18878
134 LeCun Y, Chopra S, Hadsell R, et al. A tutorial on energy-based learning. In: Predicting Structured Data. Cambridge: The MIT Press, 2006
135 Sundararaman M N, Ahmad Z, Ekbal A, et al. Unsupervised aspect-level sentiment controllable style transfer. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP), 2020. 303–312
136 Ji Y, Liu H, He B, et al. Diversified multiple instance learning for document-level multi-aspect sentiment classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 7012–7023
137 Liang B, Yin R, Gui L, et al. Jointly learning aspect-focused and inter-aspect relations with graph convolutional networks for aspect sentiment analysis. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING), 2020. 150–161
138 Ye J, Zhou J, Tian J, et al. Sentiment-aware multimodal pre-training for multimodal sentiment analysis. Knowl-Based Syst, 2022, 258: 110021
139 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2022. 27730–27744
140 Wang Y, Kordi Y, Mishra S, et al. Self-instruct: aligning language models with self-generated instructions. 2022. ArXiv:2212.10560
141 Dai W, Li J, Li D, et al. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2024
142 Ye J, Zhou J, Tian J, et al. Rethinking TMSC: an empirical study for target-oriented multimodal sentiment classification. In: Proceedings of Findings of the Association for Computational Linguistics, 2023. 270–277
143 Wei X, Zhang T, Li Y, et al. Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 10941–10950
144 Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 6077–6086
145 Riloff E, Qadir A, Surve P, et al. Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013. 704–714
146 Tay Y, Luu A T, Hui S C, et al. Reasoning with sarcasm by reading in between. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018. 1010–1020
147 Xiong T, Zhang P, Zhu H, et al. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In: Proceedings of the Web Conference (WWW), 2019. 2115–2124
148 Speer R, Chin J, Havasi C, et al. ConceptNet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2017
149 Geetha A V, Mala T, Priyanka D, et al. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. Inf Fusion, 2024, 105: 102218
150 Zhang Z, Peng L, Pang T, et al. Refashioning emotion recognition modeling: the advent of generalized large models. IEEE Trans Comput Soc Syst, 2024, 11: 6690–6704
151 Peng L, Zhang Z, Pang T, et al. Customising general large language models for specialised emotion recognition tasks. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. 11326–11330
152 Yang X, Wu W, Feng S, et al. MM-InstructEval: zero-shot evaluation of (multimodal) large language models on multimodal reasoning tasks. 2024. ArXiv:2405.07229