Sentiment Analysis: Comprehensive Reviews, Recent Advances, and Open Challenges

Qiang Lu, *Graduate Student Member, IEEE*, Xia Sun, Yunfei Long, *Member, IEEE*, Zhizezhang Gao, Jun Feng, and Tao Sun

***Abstract*— Sentiment analysis (SA) aims to understand the attitudes and views of opinion holders with computers. Previous studies have achieved significant breakthroughs and extensive applications in the past decade, such as public opinion analysis and intelligent voice service. With the rapid development of deep learning, SA based on various modalities has become a research hotspot. However, only individual modality has been analyzed separately, lacking a systematic carding of comprehensive SA methods. Meanwhile, few surveys covering the topic of multimodal SA (MSA) have been explored yet. In this article, we first take the modality as the thread to design a novel framework of SA tasks to provide researchers with a comprehensive understanding of relevant advances in SA. Then, we introduce the general workflows and recent advances of single-modal in detail, discuss the similarities and differences of single-modal SA in data processing and modeling to guide MSA, and summarize the commonly used datasets to provide guidance on data and methods for researchers according to different task types. Next, a new taxonomy is proposed to fill the research gaps in MSA, which is divided into multimodal representation learning and multimodal data fusion. The similarities and differences between these two methods and the latest advances are described in detail, such as dynamic interaction between multimodalities, and the multimodal fusion technologies are further expanded. Moreover, we explore the advanced studies on multimodal alignment, chatbots, and Chat Generative Pre-trained Transformer (ChatGPT) in SA. Finally, we discuss the open research challenges of MSA and provide four potential aspects to improve future works, such as cross-modal contrastive learning and multimodal pretraining models.**

***Index Terms*— Multimodal data fusion, multimodal representation learning, multimodal, sentiment analysis (SA), single-modal.**

I. Introduction

SENTIMENT analysis (SA) is an important yet challenging task in artificial intelligence (AI), and it aims to understand the attitudes and views of these opinion holders with computers [1]. For most time in this review, we use these three terms interchangeably. Emotion, sentiment, and affection are often involved in SA. Emotion, sentiment, and affection are often involved in SA. Emotion refers to a short-lived and intense affective response that is context-specific and involves subjective experience and goals. Sentiment refers to a persistent and stable sentiment response that is profound in experience and involves affective polarity and objects. Affection refers to a feeling of predilection, warmth, or closeness toward someone or something. Especially, in the field of SA, emotion is often described as discrete and more fine-grained emotional categories, such as joy, sadness, surprise, anger, disgust, and fear. In contrast, sentiment is represented as a more coarse-grained representation, often described as sentiment polarity, i.e., positive, negative, and neutral [2]. As an interdisciplinary research field, SA has been widely applied in daily life, such as public opinion supervision, esthetic analysis, and telephone service.

SA can be divided into single-modal SA (SSA) and multimodal SA (MSA). SSA refers to the analysis of data with a single modality, such as text, visual, and speech. In recent years, researchers have made meaningful explorations in SSA [3], [4], [5], but there are still a series of issues that still need to be resolved. The existing research has not comprehensively summed the text SA up and lacks systematic carding for the latest technology. The similarities and differences between images and facial expressions are ignored in visual SA (VSA). In speech SA, processes such as preprocessing and feature extraction could have been explained clearly, and some practical frameworks were ignored. Therefore, it is necessary to design a novel framework to comprehensively introduce and summarize the tasks and methods of SSA.

With the development of multimedia technology and social networks, people express their views and emotions in more diverse ways in the multimedia scene. Meanwhile, human cognition does not only come from single-modal data. In real scenes, multimodal data often appear in the same scene. Furthermore, it is difficult to accurately judge the sentiment state only by text or voice in some cases, such as irony. Irony often combines neutral or positive textual content and audio expression that does not match the content to complete a negative sentiment expression. The above cases are challenging to be solved fundamentally only by a single modality, and single-modal models are easily affected by noise. Therefore, MSA has attracted considerable attention in recent years.

Manuscript received 12 December 2022; revised 8 April 2023; accepted 9 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61877050 and in part by the Program for International Science and Technology Cooperation Projects of Shaanxi Province under Grant 2021KW-63. *(Corresponding author: Xia Sun.)*

Qiang Lu, Xia Sun, Zhizezhang Gao, and Jun Feng are with the School of Information Technology, Northwest University, Xi'an 710127, China (e-mail: nwulq@stumail.nwu.edu.cn; raindy@nwu.edu.cn; 202210338@stumail.nwu.edu.cn; fengjun@nwu.edu.cn).

Yunfei Long is with the School of Computer Science and Electrical Engineering, University of Essex, CO4 3SQ Colchester, U.K. (e-mail: yl20051@essex.ac.uk).

Tao Sun is with North China Petroleum Tiancheng Industrial Group Company Ltd., Renqiu City, Hebei 062552, China (e-mail: tc_st@petrochina.com.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2023.3294810.

Digital Object Identifier 10.1109/TNNLS.2023.3294810

MSA tasks combine two or more types of modal data, such as text, image, and audio, to realize SA [6]. Previous studies have attempted to utilize multimodal learning methods, but there is a heterogeneity issue; namely, the information of different modalities exists in different subspaces. Adding more modal information to the unified model can improve the performance, but it also increases the complexity and difficulty of modeling. How to map the subinformation of different spaces to a unified semantic space and realize the complementation of different modal data has become aporias in MSA. Meanwhile, MSA is a new research field, and the existing studies lack a systematic introduction to tasks and methods. Therefore, there is an urgent need for systematic induction and summary of MSA.

There are some survey papers covering the topic of SA. In the field of text, Medhat et al. [1] summarized the text SA algorithms and various SA applications, and classified them according to their contributions to various SA technologies. However, this article was an early survey, and the latest technologies were not involved such as deep learning in fine-grained SA. Abdullah and Ahmet [7] surveyed the development of deep learning architecture in text SA. They introduced the latest technology in coarse- and fine-grained SAs in detail and covered the state-of-the-art transformer-based language models. However, they overlooked some compound tasks in aspect-based SA, such as aspect-opinion pair extraction (AOPE). In the field of visuals, Ortis et al. [4] introduced the latest methods in image-based SA in detail. However, they ignore the task of facial expression recognition (FER). In the field of speech, El Ayadi et al. [5] comprehensively introduced classification schemes and databases for speech emotion recognition (SER), but the latest technologies, such as contrastive learning (CL), were not involved. In the field of MSA, Soleymani et al. [6] reviewed recent developments in MSA in different domains. However, they lacked a clear classification system and an introduction to the latest technologies. To the best of our knowledge, this article is the only survey to cover all modal SA that contains the most comprehensive tasks and the latest technologies, such as CL.

In this article, the main goal is to provide researchers with relevant advances in SA and the inner connections among them. We first take the modality as the thread to design a novel framework of SA and introduce the latest classifiers and relevant evaluation measures. Then, the workflows, trends, and datasets of SSA are reviewed, such as text, visual, and speech to guide MSA. Second, a new taxonomy is proposed to divide MSA into bimodal and trimodal MSAs based on multimodal representation learning and data fusion. The similarities and differences, advantages and disadvantages, and up-to-date methods are discussed in detail. Meanwhile, we explore the advanced large language models (LLMs) in SA. Finally, we discuss the open research challenges and provide potential aspects to improve SA's future work.

The main contributions of this work are summarized as follows.

1) We design a comprehensive framework covering the important tasks of SA and introduce the related paradigms and evaluation measures to give researchers a comprehensive understanding of relevant advances.
2) We introduce the general workflows and recent advances of single modal in detail and discuss the similarities and differences of SSA in data processing and modeling to guide MSA.
3) We provide a new taxonomy to fill the gaps of MSA and introduce the latest methods of multimodal representation learning and data fusion.
4) We summarize the commonly used datasets in different modal SA and provided researchers with guidance on data and methods according to different task types.
5) We discuss open research challenges of MSA and provide directions for the future development of SA from four potential terms.

The rest of this article is organized as follows. As shown in Fig. A1 in the Supplementary Material, in Section II, the overall framework of SA is described, including the architecture, related classifiers, and evaluation measures. In Sections III–V, the workflow, trends, and datasets of SSA are introduced and discussed, respectively. In Section VI, we introduce the new taxonomy to divide MSA into bimodal and trimodal MSAs based on representation learning and data fusion, and expand multimodal fusion technologies and alignment methods. In Section VII, we discuss the recent advancements of chatbot-based technology in SA. In Section VIII, the open research challenges are briefly reviewed, and a discussion of the future trends is presented. Finally, the conclusions are drawn in Section IX.

II. Overall Framework of Sentiment Analysis

Sentiments come from various sources in reality, such as Taobao comments, Weibo pictures, facial expressions, and audio recordings. Therefore, accurately grasping the sentiment can significantly improve the interactive experience. This section introduces a novel SA architecture to give researchers a more comprehensive understanding. Then, we describe related classifiers to better understand the studies surveyed in the later sections. Finally, the evaluation measures of SA are explained.

A. Architecture of Sentiment Analysis

SA tasks are classified into two categories: SSA and MSA, as shown in Fig. A2 in the Supplementary Material. SSA contains three types: text, visual, and speech. Text SA aims to analyze the subjective text with sentiment color to judge the sentiment polarity, which can be classified into document-level, sentence-level, and aspect-based categories. VSA establishes the relationship between image features and sentiment features, and infers the sentiment expressed by the image according to the sentiment polarity to achieve classification. VSA tasks are classified into image SA and FER. Speech SA realizes sentiment classification by modeling the linguistic and paralinguistic features of speech. Compared with the text and visual fields, SA research in the speech field belongs to an emerging field, mainly focusing on SER. MSA is divided into two types: bimodal and trimodal SAs. Bimodal SA combines two modalities, such as text and image, and trimodal SA

contains three modalities, such as text, video, and audio. Based on bimodal and trimodal types, MSA is divided into multimodal representation learning and data fusion according to a new taxonomy.

B. Related Classifiers

Multiple classifiers have been utilized for SA, such as a support vector machine (SVM), a convolutional neural network (CNN), a recurrent neural network (RNN), a recursive neural network (RecNN), and a memory network (MN) [7], but determining which works best is challenging. This section summarizes and introduces the up-to-date classifiers utilized in SA. For each architecture, a description of their operating principle is provided to understand better the studies surveyed in the later sections of this article.

1) Pretraining Model: The pretraining model (PTM) can effectively obtain knowledge from many unlabeled data and encode knowledge in parameters [8]. Early PTMs mainly focus on transfer learning, which aims to acquire important knowledge from multiple source tasks and apply the knowledge to target tasks. Recent PTMs have achieved fruitful results due to the integration of self-supervised learning and transformer [9].

2) Contrastive Learning Model: CL is discriminative self-supervised learning in self-supervised learning [10]. It is required to learn a representation learning model by automatically constructing similar and dissimilar instances. Similar instances are closer in the projection space, while dissimilar instances are farther away in the projection space via this model. The typical paradigms of CL are the agent task and the objective function. The agent task defines the positive and negative samples of comparative learning and then uses the objective function to calculate the loss to guide the learning direction of the model. The general loss function of comparative learning is defined as follows:

$$L_{i,j} = -\log \frac{\exp\left(\text{score}\left(f(x_i), f\left(x_i^+\right)\right)\right)}{\sum_{j=0}^{N} \exp\left(\text{score}\left(f(x_i), f\left(x_j\right)\right)\right)} \tag{1}$$

where score() is a function to measure the similarity between positive and negative samples.

C. Evaluation Measures

An SA task is usually modeled as a classification problem. In addition, some works also utilize regression models for SA. The commonly used evaluation measures of SA include precision (P), recall (R), accuracy (Acc), F1-score (including macro-F1 and micro-F1), mean absolute error (MAE), and correlation coefficient (Corr).

1) Precision and Recall: Precision is the proportion of correct predictions in all predictions with positive labels. Recall is the proportion of correct predictions among all positive instances.

2) Accuracy: Accuracy is the most basic evaluation measure of classification. It is the ratio of all true positive samples and true negative samples to all samples.

3) F1-Score: F1-score value comprehensively considers the factors of precision and recall, which is the harmonic function of them.

4) Mean Absolute Error: MAE refers to the average value of the distance between the predicted value $f(x)$ of the model and the true value y of the sample.

5) Correlation Coefficient: Corr is used to measure the relationship (linear correlation) between the variables x and y.

III. Text Sentiment Analysis

Text SA, also known as opinion mining, refers to the mining, analysis, and reasoning of opinions and attitudes on subjective texts with sentiment colors [3]. The rapid development of internet technology has brought people into the information and digital era, which brings convenience and a large amount of text data containing rich sentiment information. SA of text data can help the government monitor the development of public opinion and promote the harmonious development of society. Therefore, text-based SA has been widely studied and applied in academia and industry.

A. Text-Based Workflow

As shown in Fig. 1, before the text is input to the classifier, it is necessary to preprocess the raw text and convert it to word embedding. Preprocessing is a key part of text SA due to the training results of the model depend on the quality of preprocessed data. There are the following transformations: word segmentation, part-of-speech tagging, and data enhancement, which process random text data into a structured data format that can be analyzed.

Word embedding is a general term for language model and representation learning technology in natural language processing (NLP). It embeds each word from a high-dimensional space into a low-dimensional vector space, represented in a low-dimensional area by a vector on the real number field. One-hot coding is the most basic word vector that is essentially the representation of classification variables as binary vectors, but it has the problems of semantic gap and dimension disaster. To solve this problem, distributed word vectors based on the neural network are widely used in SA. After preprocessing and word embedding transformation, the word embedding vector is input to different classifiers. Then, the forward propagation process transfers the word embedding vector through network parameters to calculate the loss. It is then used in backpropagation to update network parameters and normalize them through the softmax function for classification.

B. Trends in Text Sentiment Analysis

Most studies focus on coarse- and fine-grained SAs, including document-level, sentence-level, and aspect-based, as shown in Fig. A2 in the Supplementary Material. In this section, we will detail the related tasks and development trends in three kinds of text SA.

1) Document-Level Sentiment Analysis: Document-level SA determines whether the document conveys overall positive, negative, or neutral opinions. In this case, it is a ternary classification task. It can also be expressed as a regression task, for example, inferring the overall score from one to five stars. The challenge of document-level SA is to capture sentence

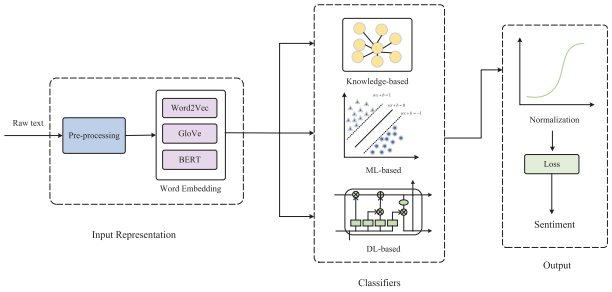

Fig. 1. Workflow of text SA.

semantic and contextual relations. Previous studies applied various methods to model document-level SA to address this challenge. The trends of document-level SA are the following parts.

a) Knowledge-based: The essence of knowledge-based models is the sentiment dictionary and grammar rules. This method uses a weighting algorithm to give sentiment vocabulary weight to build a sentiment dictionary and then uses specific calculation formula to calculate the sentiment score of sentences for classification. Hu and Liu [11] used NLP methods to identify adjectives. Then, a semantic word network WordNet [12] was built to determine the semantic direction of each adjective, and an effective algorithm was proposed to determine the opinion direction of each sentence. Due to the need for a general and complete sentiment dictionary in Chinese microblog SA, Zhang et al. [13] proposed a Chinese microblog SA method based on a comprehensive sentiment dictionary.

b) SVM-based: The SVM-based methods are usually incorporated with heavy feature engineering. SVMs separate data vectors belonging to different categories by constructing a hyperplane as a decision boundary. Pang et al. [14] applied SVM to text SA for the first time and analyzed the problem of whether the recognition sentence conforms to the theme characteristics in detail. Mullen and Collier [15] constructed feature space by using semantic direction values from different sources, and sentiment prediction was carried out by SVM. Huq et al. [16] proposed a method that used K-nearest neighbor (KNN) and SVM to analyze textual sentiment.

c) RNN-based: knowledge- and SVM-based methods have limitations such as relying on manual annotation and intensive labor, meanwhile having poor adaptability and generalization ability. RNNs transform data into distributed representation through word embedding technology and automatically learn the potential features and rules of large-scale data samples through hidden layers, so as to get rid of complex feature engineering. Xu et al. [17] introduced a caching mechanism to diversify the internal memory into several different groups with different memory cycles. In order to solve the problem that document-level SA does not consider the influence of users expressing sentiment and the evaluated products, Dou [18] proposed a deep MN for document-level sentiment classification. Because the previous method tends to assign an equally smaller weight to each word, the keywords are covered by nonsentiment words. Zhang et al. [19] proposed a cyclic attention LSTM to iteratively locate the attention region covering key sentiment words, gradually reducing the attention range and the number of tags to use the weight of key sentiment words for final sentiment classification.

2) Sentence-Level Sentiment Analysis: Unlike document-level SA to judge the overall sentiment of all sentences, sentence-level aims to determine the sentiment polarities in a single sentence. It also can be expressed as a ternary classification task. The challenge of sentence-level SA is to model the semantic relations of all words in a sentence and capture the syntactic dependencies. Relevant studies are given as follows.

a) CNN-based: The local receptive field and weight-sharing operation of CNN can extract the local features of text well. Kalchbrenner et al. [20] provided a dynamic CNN (DCNN) to model the semantic of sentences. DCNN used dynamic k-max pooling to capture short- and long-term relationships and generate a feature graph on sentences. To extract information from sentences in a more standardized way, Dos Santos and Gatti [21] provided a character-to-sentence CNN (CharSCNN). They used two convolution layers to extract relevant features from words and sentences of any size. However, different types of sentences express sentiment differently, while the traditional models only focus on certain sentence types. To address these problems, Chen et al. [22] provided a novel framework based on CNN for sentence-level SA. They used a divide-and-conquer approach to deal with different types, which included nontarget, one-target, and multitarget sentences.

b) CNN-LSTM-based: The CNN-LSTM is a class of architectures combining CNN and LSTM. CNN can extract local information but may fail to capture long-distance dependencies. LSTM can solve this limitation by modeling sentence sequences. Wang et al. [23] provided a joint CNN and RNN architecture for sentiment classification. Wang et al. [24] proposed a regional CNN-LSTM model to predict sentiment. Based on [24], Wang et al. [25] provided a region division strategy to improve the performance of SA.

3) Aspect-Based Sentiment Analysis: Coarse-grained SA aims to identify the overall sentiment toward the whole document or sentence, ignoring the problem that there may be multiple aspects. There may be multiple entities in a document or a sentence for document- and sentence-level SAs. In this case, each entity expresses different sentiment polarity. Therefore, only analyzing the overall sentiment toward the whole document or sentence will lead to inaccurate classification. ABSA has received increasing attention due to its ability to identify each specific entity in the sentence and analyze the sentiment of entities.

There are four sentiment elements of ASBA, as shown in Fig. 2: aspect term, aspect category, opinion term, and sentiment polarity. Depending on whether the output is a single element or a coupling element, ABSA tasks are classified into two categories: single ABSA and compound ABSA. The input and output of each task are shown in Table I, where S is the sentence, and T, C, O, and P represent the aspect term, aspect category, opinion term, and sentiment polarity, respectively. Single ABSA includes the following subtasks: aspect term

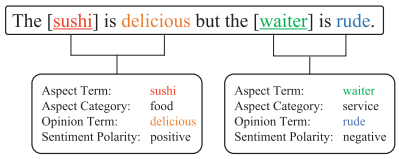

Fig. 2. Example of aspect-based SA.

extraction (ATE), aspect category detection (ACD), opinion term extraction (OTE), and aspect sentiment classification (ASC).

a) Aspect term extraction: ATE is to extract the explicit targets in a given sentence [26]. For instance, ATE aims to extract two aspect terms "sushi" and "waiter" in the sentence "The sushi is delicious but the waiter is rude." The challenges of ATE are that domain-specific knowledge and a large amount of labeled data are required. Yin et al. [27] learned the distributed representation of words and dependent paths from the text corpus, and they used word embedding, linear contextual embedding, and dependent path embedding to enhance conditional random field (CRF) to extract aspect terms. Despite the great results of the above methods, they depend on the labeled data. Giannakopoulos et al. [28] introduced an unsupervised and domain-independent method for annotating raw opinion text and provided a classifier based on B-LSTM and CRF for both unsupervised and supervised ATEs. Venugopalan and Gupta [29] proposed a semantic filter based on BERT that combined semantic information to enhance co-occurrence statistics.

b) Aspect category detection: ACD aims to infer the categories of aspects mentioned in the sentence [30], for example, in the sentence "The sushi is delicious but the waiter is rude." Based on the given entities "sushi" and "waiter," the aspect categories "food" and "service" are concluded. Schouten et al. [31] presented an unsupervised and supervised method that could find the category of aspects according to the co-occurrence frequencies. To address the problem of failing to recognize aspect categories that only contained a few labels, Hu et al. [32] proposed a multilabel few-shot learning (FSL) method based on the prototypical network. They alleviated the noise by aspectwise attention and query-set attention. Similar to [32], Liu et al. [33] applied a multilabel FSL method to meet the challenges of aspect sharing, aspect interference, and aspect diversity, and proposed a novel label-enhanced prototypical network (LPN) for ACD.

c) Opinion term extraction: OTE is used to identify and extract opinion terms toward the related aspect [34]. Due to the fact that opinion terms and aspect terms always appear together, OTE is also referred to as target-oriented opinion word extraction (TOWE). Fan et al. [34] proposed a novel sequence labeling subtask for ABSA that aimed at extracting the corresponding opinion words for a given opinion target. Veyseh et al. [35] introduced a novel regularization technique and leveraged the syntax-based opinion possibility scores and the syntactic connections between the words. Mensah et al. [36] adapted a GCN model to enhance word representations to examine the actual contribution of each component in TOWE.

TABLE I
INPUT AND OUTPUT OF EACH TASK

Tasks	Input	Output
Aspect Term Extraction (ATE)	S	T
Aspect Category Detection (ACD)	S	C
Aspect Sentiment Classification (ASC)	S, T	P
Opinion Term Extraction (OTE)	S, T	O
Aspect-Opinion Pair Extraction (AOPE)	S	$\langle T, O \rangle$
Aspect Sentiment Quad Prediction (ASQP)	S	$\langle T, C, O, P \rangle$
Aspect Sentiment Triplet Extraction (ASTE)	S	$\langle T, O, P \rangle$
E2E Aspect-based Sentiment Analysis (E2E ABSA)	S	$\langle T, P \rangle$

d) Aspect sentiment classification: ASC, as known as aspect-based sentiment classification, is to predict the sentiment polarities of each aspect in a given sentence. ASC can be divided into aspect-category and aspect-term sentiment classification. The aspect category sentiment classification implicitly describes the general entity category, and the aspect term sentiment classification characterizes specific entities that occur explicitly in a sentence. With the deepening of deep learning research, ASC methods have gradually transitioned to deep learning method systems. Deep learning methods transform data into distributed representation through word embedding technology and automatically learn the potential features and rules of large-scale data samples through hidden layers, so as to get rid of complex feature engineering. The ASC methods are divided into CNN-based, LSTM-based, MN, GCN-based, and PTM-based, and the analysis and comparison of different classifiers are shown in Table II.

CNN-Based: Huang and Carley [37] proposed a parameterized filter CNN and a parameterized gate CNN for ASC. As the CNN-based methods are difficult to use important aspects of location information in a unified framework, Wang et al. [38] proposed a unified position-aware CNN (UP-CNN) that generated position embedding according to the relative distance between each word and a given aspect.

RNN-Based: Despite that CNN-based methods could capture local semantic information, pooling operations resulted in the loss of overall semantic dependence. Tang et al. [39] first proposed two LSTM models that automatically capture target information. To address the problem that LSTM-based methods introduced noise in the process of feature selection and extraction, Liang et al. [40] proposed an aspect-guided gated recurrent unit (GRU) encoder to guide sentence coding and force the model to use the generated sentence representation, which reconstructed the given aspect. The attention mechanism breaks the limitation of LSTM that the input depends on the output of the previous time. Therefore, Wang et al. [41] first introduced the attention mechanism into aspect-based sentiment classification. Most existing methods ignored the role of position information; Gu et al. [42] proposed a position-aware bidirectional attention network (PBAN) based on GRU that converted the position information into position embedding.

RecNN-Based: RecNN is an effective extension of RNN, which is a kind of neural network with a tree structure and recurses through the nodes of the tree structure.

TABLE II
ANALYSIS AND COMPARISON OF DIFFERENT ASPECT-BASED SENTIMENT CLASSIFICATION METHODS

Structures	Advantages	Disadvantages	Models
CNN-based	Captures local semantic information, and weight sharing.	Fails to capture long-distance semantic dependency.	[37], [38]
RNN-based	Suitable for processing sequence data, taking into account historical information.	Limited ability of memory, vanishing gradient and exploding gradient.	[39], [40], [41], [42]
RecNN-based	Good ability to handle tree and graph structures.	Weak memory storage capacity.	[43], [44], [45]
MN-based	Introduces external storage to remember relevant information	Low memory capacity and poor fault tolerance	[46], [47], [48]
GCN-based	Strong performance in capturing syntactic dependencies.	Poor expansibility and high complexity.	[49], [50], [51]
PTM-based	Beneficial to downstream tasks via the pre-learned knowledge, better generalization performance and fast convergence.	Needs massive data support, and huge consumption of computer resources.	[52], [53], [54], [55]

Dong et al. [43] proposed an adaptive RecNN for aspect-level SA on Twitter. Nguyen and Shirai [44] proposed a phrase RecNN to make the representation of the target aspect richer by using syntactic information from both the dependence and constituent trees of the sentence. To address the limitation that previous methods rely on hand-coded rules, Wang et al. [45] proposed a novel joint model that integrates RecNNs and CRFs into a unified framework for explicit aspect and opinion terms coextraction.

MN-Based: MN uses memory components to store information for long-term memory functions. Tang et al. [46] introduced a deep MN for aspect-level sentiment classification. Majumder et al. [47] presented a novel method of incorporating the neighboring aspect-related information into the sentiment classification of the target aspect using MNs. As previous models still face the issues of the weakness of pretrained word embeddings and weak interaction between the specific aspect and the context in attention mechanism, Liu and Shen [48] proposed a novel end-to-end memory neural network (ReMenNN) that contained an embedding adjustment learning module and a multielement attention mechanism.

GCN-Based: CNN, RNN, and an attention mechanism show excellent performance in capturing semantic information, but ignoring an important problem, i.e., syntactic dependence, and may mistakenly use context-free information as clues for identifying sentiment. The graph convolutional network (GCN) performs a convolution on the top of the LSTM in the form of an L-layer to create context-aware nodes, and the hidden representation of each node is updated through a graph convolution operation with a normalization factor [49]. Zhang et al. [50] applied GCN to aspect-based SA for the first time and proposed a novel aspect-based sentiment classification framework. Li et al. [51] proposed a dual graph convolution network (DualGCN). They utilized the probability matrix from the dependence parser to build the syntax-based GCN (SynGCN) and then used the self-attention mechanism to build the semantic-based GCN (SemGCN) for ASC.

PTM-Based: Hoang et al. [52] proposed a combination module that utilized the BERT to generate context word representation to classify aspects and sentiment. The existing PTM only takes the pretrained BERT as a black box, which lacks context awareness. Wu and Ong [53] proposed context-guided BERT (CGBERT) and quasi-attention CGBERT (QACG-BERT) for ASC. In order to capture reasonable attention weight, Wang et al. [54] provided the intralevel and interlevel attention mechanisms based on BERT to generate the hidden representation of a sentence and constructed a focus attention mechanism to enhance sentiment identification. No BERT model currently considers topic information. Zhou et al. [55] developed two variants of TopicBERT. TopicBERT-ATP captured topic information through auxiliary training tasks, and TopicBERT-TA achieves sentiment classification by dynamically changing topics.

Compound ABSA is to extract multiple elements and couple them in different tasks and can be divided into the following tasks: AOPE, aspect sentiment triplet extraction (ASTE), aspect sentiment quad prediction (ASQP), and end-to-end ABSA (E2E ABSA).

e) Aspect-opinion pair extraction: AOPE is defined as extracting aspects and opinion expressions along with their relations [56]. For example, in Fig. 2, AOPE is to extract the pairs ⟨sushi, delicious⟩, ⟨waiter, rude⟩. There are two ways: one is first to extract aspect terms and opinion terms and then pair them; the other is first to perform ATE and then identify the corresponding opinion terms for each predicted aspect term. The challenge of AOPE is that ATE and OTE are interrelated and mutually reinforcing. Chen et al. [57] proposed a synchronous double-channel recurrent network (SDRN) for AOPE. Based on the second way, Gao et al. [58] designed a question-driven span labeling (QDSL) model to extract aspect-opinion pairs.

f) Aspect sentiment triplet extraction: ASTE aims to discuss relations of the sentiment elements that, what aspect term is, how is the sentiment polarity and why is this sentiment expressed [59]? It is similar to AOPE, which outputs the tripe ⟨aspect term, opinion term, sentiment polarity⟩. For example, in the sentence in Fig. 2, ASTE aims to extract sentiment tripe ⟨sushi, delicious, POS⟩, ⟨waiter, rude, NEG⟩. Compared with a single ABSA, ASTE contains more abundant sentiment information to indicate sentiment elements and their relations. The challenges of ASTE are similar to AOPE that also contains corresponding relations between three elements. Peng et al. [60] first introduced an ASTE task and proposed a two-stage framework to address this task. To address the aforementioned challenges, Chen et al. [61] proposed a bidirectional MRC framework to formalize the ASTE task as

a machine reading comprehension (MRC) task. The excellent performance of the PTMs has been verified, but it is inefficiency due to large-scale parameters. Zhang et al. [62] construct a structural adapter, triplet parser, and triplet decoder to uncover continuous tokens and generate aspect sentiment triplets.

g) Aspect sentiment quad prediction: ASQP is used to extract quadruples of aspect term, aspect category, opinion term, and sentiment polarity [63]. For example, in Fig. 2, ASQP is to extract the all elements ⟨sushi, food, delicious, POS⟩, ⟨waiter, services, rude, NEG⟩. Zhang et al. [63] introduced the ASQP task and proposed a novel paraphrase modeling paradigm to cast the ASQP task to a paraphrase generation process. Bao et al. [64] proposed a pretrained model to integrate both syntax and semantic features to jointly detect all sentiment elements in a tree. Gao et al. [65] proposed a unified generative multitask framework to solve multiple ABSA tasks by controlling the type of task prompts consisting of multiple element prompts.

h) End-to-end aspect-based sentiment analysis: End-to-end aspect-based SA is designed to extract aspect term and its corresponding sentiment polarity simultaneously [66]. In the sentence in Fig. 2, E2E-ABSA aims to extract pairs ⟨sushi, POS⟩, ⟨waiter, NEG⟩, and it can be divided into two subtasks: ATE and ASC. Li et al. [67] first presented a unified end-to-end model to solve the complete tasks of ABSA. To address the imbalance of labels of E2E-ABSA, Luo et al. [68] proposed a GRadient hArmonized and CascadEd labeling model (GRACE) to capture the interaction between aspect terms with a stacked multiattention module for SA.

C. Datasets and Performance Summary of Text Sentiment Analysis

In text SA, datasets perform essential roles in achieving an excellent performance of models. This section details common datasets for document-level, sentence-level, and aspect-based SA, and shows the performance summary of text SA. An overview of these datasets is shown in Tables A1 and A5 in the Supplementary Material.

IV. Visual Sentiment Analysis

The development of social media has brought new challenges to SA [4]. On the basis of expressing opinions through words, more and more people tend to use images and videos to describe their experiences and express sentiment. The information contained in visual content is not only related to the semantic content, such as the obtained objects or actions, but also related to the sentiment conveyed by the depicted scenes. Therefore, VSA is very important to understand the sentiment effects (i.e., induced emotions) beyond semantics. Because the primary visual features and sentiment semantic features of images exist in unequal subspaces, the task of VSA is very challenging.

A. Visual-Based Workflow

As shown in Fig. 3, the processing of VSA can be divided into the following steps: preprocessing, feature extraction, classifier design, and sentiment classification. Due to the original image data's differences in size, color, and space, it needs to be preprocessed by graying processing [69], geometric transformation [70], and image enhancement [71]. First, the processing speed of the model is improved by graying processing. Standard gray processing methods include component, maximum, average, and weighted average methods. Then, the geometric transformation processes the image through translation, transposition, and scaling to correct the error. Finally, image enhancement is aimed at applying a given image and purposefully emphasizes the global or regional features of the image, mainly including spatial and spectral methods.

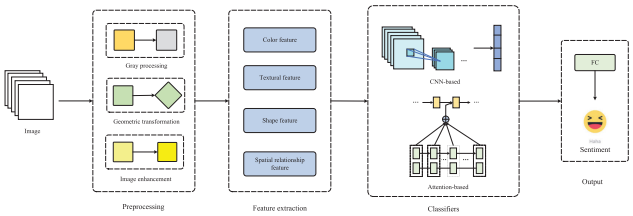

Fig. 3. Workflow of VSA.

Feature extraction is the critical step of VSA, focusing on extracting the visual features related to image sentiment. There are four visual features in general: color, textural, shape, and spatial relation. The color feature is a global feature that describes the surface properties of the scene corresponding to the image or image region. The textural feature is also a global feature, but it only describes the features of the object's surface. There are two kinds of representation methods for shape features: one is the profile feature and the other is the region feature. The profile features mainly aim at the outer boundary of an object, while the regional features relate to the whole shape region. Finally, spatial relationship features refer to the mutual spatial position or relative direction relationship between multiple objects segmented from the image. After feature extraction, image features are input into classifiers to generate a hidden representation for sentiment classification.

B. Trends in Visual Sentiment Analysis

VSA tasks mainly focus on images and facial expressions. The purpose of them is to recognize and analyze the sentiment expressed but have some differences. Image SA is a visual analysis of nonverbal sentiment expression in social media, which is aimed at analyzing the sentiment of publishers or observers. FER extracts facial expressions or body postures from individuals or groups to judge emotions. Therefore, VSA is classified as image SA and FER according to attributes and objects, as shown in Fig. A2 in the Supplementary Material, and analysis and comparison of different methods are shown in Table III. FER is divided into a static image method and a dynamic image method according to different data types.

1) Image Sentiment Analysis: Image SA extracts and combines the global or regional features of the image and classifies the sentiment by establishing a relationship with the sentiment semantics. The artificial method relies on manually extracting

the primary features of the image, such as color, textural, and profile features, for sentiment classification [72], [73], [74].

a) CNN-based: The semantic gap between image features and sentiment features is solved with feature-based methods to some extent, but these methods rely on low-level features. With the development of deep learning technology, CNN models can automatically extract medium- and high-level features from datasets that are widely used in image SA. You et al. [75] used large-scale weak-label training data for learning, and then, they fine-tuned the CNN by using the progressive training strategy and the domain transfer strategy to classify sentiment. Wang et al. [76] proposed a novel VSA approach with deeply coupled adjective and noun neural networks to address three challenges: large intraclass variance, fine-grained image categories, and scalability. Yang et al. [77] proposed a weakly supervised coupled network that integrated visual sentiment classification and detection into a unified CNN framework. CNN is often used in conjunction with attention mechanisms; You et al. [78] first considered the impact of regional image areas on VSA and used attention mechanisms to match local image areas with descriptive visual attributes. As image SA can be specified as the gradual perception of image regions from semantics to sentiment, Zhang et al. [79] thought that the mining of sentiment-related regions was of great significance for sentiment recognition and proposed a multilevel sentiment region correlation analysis model.

2) Facial Expression Recognition: FER is to capture facial expression and feature to achieve sentiment classification. As facial expression change is a complex process, involving muscle movement, psychological, and environmental factors, existing studies usually only consider the changes of facial shape and texture caused by facial muscle movement.

FER includes four parts: face image extraction, face detection, feature extraction, and feature classification. The technology of face image extraction and face detection has been very mature, so the research methods focus on feature extraction and classification. There are two states of face image: static face image and dynamic face image. The static face image has locality and timeliness, and presents the expression state of a single image when the expression occurs. The dynamic face image has integrity and activity, and presents the movement process of expression between multiple images. Therefore, FER can be divided into the static face image method and the dynamic face image method according to different data types.

a) Static face image: Considering the influence of face changes on global information, Shu-Ren et al. [80] proposed a FastICA algorithm, which combined the hidden Markov model (HMM) for expression recognition. Zhang et al. [81] extracted SIFT features for facial expression classification, which corresponds to a group of landmarks from each facial image.

CNN-Based: With the development of deep learning, FER attempts to capture high-level abstraction features through neural network architectures of multiple nonlinear transformations and representations. Pons and Masip [85] proposed a method of weighting CNN classifiers, which used CNN to learn the nonlinear relationship between classifiers to better distinguish basic sentiment. Inspired by visual attention described in cognitive neuroscience, Farzaneh and Qi [86] designed a depth measurement learning method based on modular attention for FER.

GAN-Based: However, CNNs rely on a large number of labeled data; especially, image annotation is time-consuming and laborious. The generative adversarial network (GAN) [96] is a kind of unsupervised learning model, which produced good outputs through mutual game learning of the generative model and the discriminative model. Lai and Lai [89] proposed a multiview FER method for multitask learning with GANs to predict the expression class label of the input face. Zhang et al. [90] proposed an end-to-end deep learning model, which combined different gestures and expressions to perform FER with unchanged posture. Cai et al. [91] proposed a novel Identity-Free conditional GAN (IF-GAN) for FER.

CL-Based: Most of the above methods are based on supervised training, but the annotation data are always limited. PTMs have proved that self-supervised pretraining can learn prior knowledge distribution from a large number of unlabeled data, and excellent results can be obtained through fine-tuning downstream tasks. In recent years, the research focus of FER has shifted from supervised pretraining to self-supervised pretraining, and comparative learning is an important support [97], [98]. CL is discriminative self-supervised learning in self-supervised learning. It is required to learn a representation learning model by automatically constructing similar and dissimilar instances. Similar instances are closer in the projection space, while dissimilar instances are farther away in the projection space via this model. Shu et al. [92] proposed an effective self-supervised CL framework for FER. In view of two concerns that arousal-valence-based FER approaches have not yet dealt with: the key for feature learning of facial emotions and the facial emotion-aware features extraction, Kim and Song [93] incorporated visual perception ability into representation learning for the first time to focus on semantic regions that are important for emotion representation.

b) Dynamic face image: Despite good results of the method based on static facial images, it fails to consider time information and subtle appearance changes that are not available in real-world scenes. The dynamic face image method reflects the process of facial expression change over a period of time. It takes a series of frames in the time window as input and uses textural and time information to encode subtle expressions. Traditional methods can be divided into optical flow methods, model methods, and geometric methods. The optical flow method is used for moving object detection, which uses the change of pixels in the time domain and the correlation between adjacent frames to find the corresponding relationships [82], [83], [84].

CNN-Based: Deep learning networks are designed to encode temporal dependencies in consecutive frames and have been shown to benefit from learning spatial features in conjunction with temporal features. Jung et al. [87] used a limited number of image data to identify facial expressions to overcome the problem of small amounts of data. In view of the deep-level image extraction capability of CNN and the time-series data

TABLE III
ANALYSIS AND COMPARISON OF DIFFERENT VSA METHODS

Structures	Advantages	Disadvantages	Models
Future-based	Good performance of small sample data and has good relevance.	Time-consuming and labor-intensive, limited learning ability.	[72], [73], [74], [80], [81], [82], [83], [84]
CNN-based	Local receptive field and weight sharing.	Redundant and inefficient, unable to understand advanced features.	[75] , [76], [77], [78], [79], [85], [86], [87], [88]
GAN-based	Training by unsupervised learning, and adopts mutual game learning to product good outputs.	Not suitable for processing discrete data and difficult to train.	[89], [90], [91]
CL-based	Learns prior knowledge from a large number of unlabeled data via self-supervised learning.	Bias of training data and hard negative example concerns.	[92], [93]
Transformer-based	Global characteristics and good modal representation capability.	Weak ability to capture local information, and excessive demand for computing power.	[94], [95]

processing capability of LSTM, Donahue et al. [88] proposed a long-term recurrent convolutional network (LRCN) with dual depth in space and time.

Transformer-Based: In recent years, the success of transformers inspired researchers to use transformer encoders in FER. Zhao and Liu [94] proposed a dynamic FER transformer (Former-DFER) for the in-the-wild scenario. They designed a convolutional spatial transformer (CS-Former) and a temporal transformer (T-Former) to learn more discriminative facial features and deal with the issues such as occlusion, nonfrontal pose, and head motion. Li et al. [95] thought that the above method [94] ignored distinguishing the key frames and the noisy frames; they proposed a noise-robust dynamic FER network (NR-DFERNet) to reduce the interference of these noisy frames.

C. Datasets and Performance Summary of Visual Sentiment Analysis

Training data are important for VSA. In this section, we describe the publicly available datasets that contain a large number of affective images in VSA and show the performance summary of VSA, as shown in Tables A2 and A7 in the Supplementary Material.

V. SPEECH SENTIMENT ANALYSIS

As the main medium in daily communication, speech contains abundant paralinguistic features in the transmission process, such as sentiment, purpose, and state [99]. However, language understanding is a very complex process, and human speech emotion change is an abstract dynamic process, which is difficult to describe its emotional interaction with static information. How to model the linguistic and paralinguistic features of speech to understand the meaning of speech is a challenging task.

Speech SA is known as SER. It is a computer simulation of the above sentiment perception and understanding process of humans [100]. The computer is used to analyze emotions, extract emotion features, and use parameters to conduct corresponding modeling and recognition, Then, the mapping relations between features and emotions are established to classify emotions.

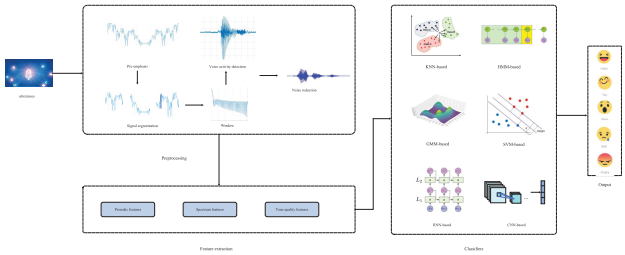

Fig. 4. Workflow of speech SA.

A. Speech-Based Workflow

The workflow of speech SA focuses on the following parts: speech processing, speech feature extraction, and sentiment classification. Overall processing of speech SA is shown in Fig. 4.

1) Speech Processing: Speech processing aims to automatically suppress interference signals and consists of the following steps: preemphasis [101], speech segmentation [102], windowing [103], voice activity detection [104], and noise reduction [105].

2) Speech Feature Extraction: The speech emotion features can be divided into linguistic features and acoustic features. Language features are the speech information that voice expresses; acoustic features include the speaker's tone, intonation, and emotion color. Extracting acoustic features with high correlation is helpful to determine the speaker's emotional state. Generally, acoustic features are extracted on a frame basis, but these features are generally used as the input of the model to perform emotion recognition in a global statistical way. At present, the commonly used acoustic features include prosodic features, spectrum features, and tone quality features.

3) Speech Classifiers: Classifiers for speech SA include two categories: traditional machine learning methods and deep learning methods. Numerous classifiers have been utilized for the SA, but determining which works best is difficult. Therefore, the ongoing researches are widely pragmatic. We present the analysis and a comparison of different classifiers in Table IV.

B. Trends in Speech Sentiment Analysis

Speech SA has been initially explored in the past 20 years, and features based on manual extraction have been widely used in it. However, human speech emotion change is an abstract dynamic process, which is difficult to describe its emotional interaction with static information. The rise of machine learning brings new opportunities for the development of SER. Traditional machine learning methods are used in speech SA because of their fast speed and high interpretability. These methods recognize the respective classes and samples by approximating the mapping function for classification and can be divided into KNN, HMM, the Gaussian mixture model (GMM), and SVM.

1) KNN-Based: KNN is a supervised learning algorithm, whose essence is to calculate the distance between different eigenvalues to classify samples [122]. The method of Feraru and Zbancioc [106] proposed an improved version of the KNN algorithm, which was associated with each parameter for SER according to the performance of feature vector weight in the classification processing. Rieger et al. [107] utilized the integration of pattern recognition paradigm with spectral feature extraction (including CEP, MFCC, LSF, ACW, and PFL) and KNN classifiers to perform SER.

2) HMM-Based: HMM is a statistical model used to describe a Markov process with hidden unknown parameters [123]. Nwe et al. [108] proposed a text-independent speech emotion classification method based on HMM, which used LFPC to represent speech signal and HMM as a classifier. Schuller et al. [109] introduced the time complexity into HMM and considered the low-level instantaneous features rather than the multiple states of global statistics.

3) GMM-Based: GMM divides objects into several Gaussian probability density functions to accurately quantify objects [124]. Ayadi et al. [110] proposed a Gaussian mixture vector autoregressive model, which modeled the dependence between extracted speech feature vectors and the multimodality in their distribution. Mishra and Sekhar [111] discussed the applicability of the variational methods based on the parameters of GMM to SER.

4) SVM-Based: Seehapoch and Wongthanavasu [112] provided an SVM method to recognize and classify the speech emotion from Berlin, Japan, and Thai emotion datasets. In order to solve the problem of emotional confusion in multi-SER, Sun et al. [113] proposed an SER method based on the decision tree SVM model with Fisher feature selection. Jain et al. [114] proposed SVM methods based on a one-against-all (OAA) strategy and a gender-dependent strategy for sentiment emotion recognition.

Neural networks, with their characteristics of nonlinear mapping, generalization, and fault-tolerant, make this kind of method have both good real-time performance and high recognition accuracy. The deep learning method-based speech emotion analysis can be divided into CNN-based, RNN-based, and PTM-based.

5) RNN-Based: Lee and Tashev [115] proposed a learning method with Bi-LSTM, which could extract a high-level representation of the emotional state in the temporal dynamics. Because the attention machine can focus on important information in the sequence, Mirsamadi et al. [116] combined Bi-LSTM with a novel pooling strategy for SER.

6) CNN-Based: Liu et al. [117] proposed a feature fusion method based on CNN, which combined spectral features and hyperprosody features to classify speech emotions. Neumann and Vu [118] integrated the representation learned by an unsupervised automatic encoder into a CNN emotion classifier and used unsupervised representation learning to improve the performance of SER. To address the issue of lacking real-time speech processing, Kwon et al. [119] proposed an E2E real-time model based on a 1-D dilated CNN (DCNN).

7) PTM-Based: Li et al. [120] proposed a contrastive predictive coding (CPC) for SER. This method contains two stages. First, a feature extractor model with CPC on a large unlabeled dataset was pretrained. Then, an emotion recognizer with features learned in the first stage was trained for SER. To address the issues that the above models lacked in both accuracy and learning robust representations agnostic to changes in voice, Alaparthi et al. [121] proposed supervised CL with transformers for SER and verified the comparison settings through different enhancement strategies.

C. Datasets and Performance Summary of Speech Sentiment Analysis

The speech sentiment database is the database of SER, and its quality directly determines the performance of the model. In addition, considering the difference between the classification framework and tasks, the design purpose and strategy of the emotion database are very important. We collect these datasets and show the performance summary of speech SA, as shown in Tables A3 and A8 in the Supplementary Material.

VI. Multimodal Sentiment Analysis

With the internet and multimedia technology development, text, image, and voice data are growing exponentially. Multimodal data have gradually become the main form of data. Due to the limited information obtained by SSA, achieving effective analysis in some specific scenarios is difficult. Therefore, the existing research began to try to model MSA.

MSA aims to combine two or more modalities of data, such as text, image, and audio, to realize the understanding and analysis of people or topics through the information complementarity between different modal data [6]. Existing research on MSA focuses on constructing multimodal feature vectors. Although both multimodal representation learning and multimodal data fusion can obtain intermediate feature vectors, there are differences between them: multimodal representation learning aims to learn the semantic representation of modal data to be applied to downstream tasks. It is divided into joint representation and collaborative representation. On the other hand, multimodal data fusion aims to integrate multimodal data with a certain framework and methods to jointly contribute to solving the target task. It is divided into three methods: early, late, and hybrid. This section proposes a new classification method of MSA, which divides MSA into methods based on multimodal representation learning and multimodal data fusion.

TABLE IV
ANALYSIS AND COMPARISON OF DIFFERENT ALGORITHMS IN SPEECH SA

Structures	Advantages	Disadvantages	Models
KNN	High fitting ability and easy to realize.	Large amount of calculation and poor interpretability.	[106], [107]
HMM	Suitable for the identification of time series and the system has good scalability.	High model complexity, poor fitting function and robustness.	[108], [109]
GMM	Strong fitting ability and robustness.	The initial solution and order of the model are too high, and the dependence on training data is strong.	[110], [111]
SVM	High fitting ability, strong robustness and global optimization.	Large scale training samples are difficult to implement and require a large amount of memory.	[112], [113], [114]
RNN	Good sequence modeling ability and memory ability.	Gradient disappearance and gradient explosion.	[115], [116]
CNN	Weight sharing, and strong generalization ability.	Ignoring the correlation between the local and the whole, it is easy to fall into the local minimum.	[117], [118], [119]
PTM	Learns prior knowledge from a large number of data.	Requires a lot of computing resources and has high complexity.	[120], [121]

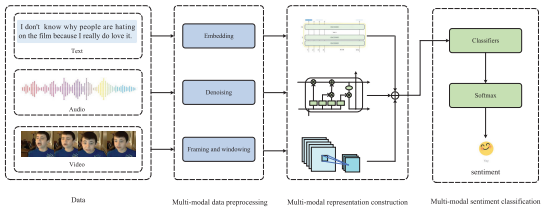

Fig. 5. Workflow of MSA.

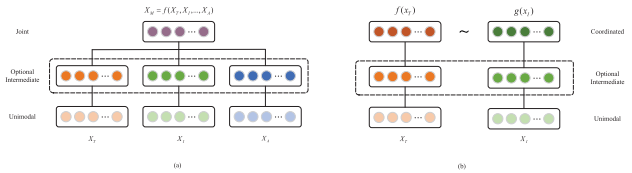

Fig. 6. Architecture of multimodal representation learning. (a) Joint representation learning. (b) Coordinated representation learning.

A. Multimodal Workflow

As shown in Fig. 5, MSA methods mainly include three steps: multimodal data preprocessing, multimodal representation construction, and multimodal sentiment classification.

1) *Multimodal Data Preprocessing:* Word segmentation and POS tagging are performed for text data, and then, words are mapped to continuous low-dimensional vector space through word2vec or GloVe. For image data, denoising is performed by smoothing technology. For audio data, framing, windowing, and Fourier transform are performed.
2) *Multimodal Representation Construction:* The preprocessed multimodal data were mapped to a unified semantic space, including multimodal representation learning and multimodal data fusion.
3) *Multimodal Sentiment Classification:* The constructed multimodal representations are input into the classifier to obtain hidden vectors and then normalize them according to different tasks to predict the polarity of sentiment or classify emotions.

B. Multimodal Representation Learning and Data Fusion Methods

1) Multimodal Representation Learning-Based: Multimodal representation learning maps different modal data to a unified semantic space so that the representation contains information across different modalities.

The difficulty lies in the heterogeneity of multimodal data and how to use the complementarity and consistency of different modal information to represent the data. Therefore, it is divided into joint representation learning and coordinated representation learning, as shown in Fig. 6.

Joint representation learning maps the information of multiple modes, such as text X_T, image X_I, and audio X_A into a unified multimodal vector space $X_M = f(X_T, X_I, \ldots, X_A)$. The unified multimodal vector aims to capture the complementarity.

Coordinated representation is responsible for mapping each mode to its own representation space, but the mapped vectors meet certain correlation constraints. For example, unlike joint representation learning, coordinated representation learning represents text X_T, image X_I, and audio X_A separately and then coordinates the relationship between different modes through constraints $f(x_T) \sim g(x_I)$.

Since coordinated representation learning preserves the information of original modes, and its optimization objective is the cooperative relationship between different modes, it is suitable for applications with only one mode as input, such as multimodal retrieval and translation. On the other hand, joint representation learning can only obtain a unified representation in the end. Its ultimate optimization goal is model prediction performance, which is suitable for applying multimodal input, such as MSA. Therefore, the research of MSA based on multimodal representation learning focuses on joint representation learning.

2) Multimodal Data Fusion-Based: With the emergence of deep learning models, the boundary between multimodal representation learning and multimodal data fusion has become blurred [125]. Multimodal data fusion predicts the results by

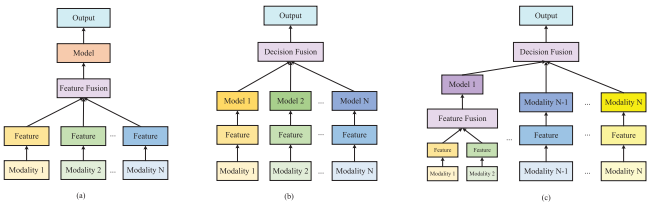

Fig. 7. Schematic diagrams of different data fusion strategies. (a) Early fusion. (b) Late fusion. (c) Hybrid fusion.

integrating information from multimodalities. Like multimodal representation learning, multimodal data fusion can also obtain intermediate feature vectors containing multimodal information. Still, there are differences between them: multimodal representation learning focuses on whether the multimodal representation has good properties, such as smoothness, sparsity, and natural clustering, and can be well applied to downstream tasks; multimodal fusion focuses on how to integrate multimodal data with a certain architecture or approach and jointly contribute to solving the target task [126]. Multimodal fusion methods are divided into early, late, and hybrid. Their typical structures are shown in Fig. 7, and the analysis and a comparison of different classifiers are shown in Table V.

a) Early fusion: Early fusion, also known as feature-level fusion, completes the fusion of features before inputting the classifier by extracting features from different modal information. Early fusion can better capture the interaction between modalities, and only one model needs to be trained to complete the feature fusion of different modalities. Therefore, it is widely used in the early research of MSA.

b) Late fusion: Late fusion is also called decision-level fusion. Different modal features are modeled separately, and then, the output from the model is integrated to produce the final prediction. The processing of late fusion is irrelevant to features and requires multinetwork models for training that can adapt well to the problem of modal missing.

c) Hybrid fusion: The hybrid fusion method combines early fusion and late fusion. Early fusion fails to make good use of the complementarity of different modalities, and late fusion has the problem of missing modalities and redundancy. Therefore, the hybrid model can capture the complementarity of modal parts and reduce the modal redundancy by combining early and late fusion. Due to the diversity and flexibility of neural networks, there are a large number of methods that adopt hybrid fusion strategies.

C. Bimodal Sentiment Analysis

Bimodal SA aims to combine data from two modalities, such as text and image to predict sentiment. In this section, we divide bimodal SA based on multimodal representation learning and data fusion methods.

1) Multimodal Representation Learning-Based Methods: Aguilar et al. [127] researched MSA from the perspective of speech and vocabulary, and proposed a method combining acoustics and vocabulary. When sentiment information is transmitted across different domains, domain-specific expressions should be deleted to reduce domain transfer of expression style. Zhang et al. [128] proposed a disentangled sentiment representation adversarial network (DiSRAN). They first employed the cross-modal attention layer to obtain the aligned multimodal joint representation with rich multimodal semantic interaction, then used the sentiment embedding module to separate the sentiment information, and alleviated the style differences via adversarial training.

2) Multimodal Data Fusion-Based Methods: In early fusion methods, Wimmer et al. [129] proposed a method to extract low-level audio and video features at a frame rate that combined video- and audio-based low-level descriptors (LLDs) to obtain a representative and robust sentiment classification feature set through functional analysis. The asynchrony of sentiment patterns and the fuzziness of different modalities make MSA more complicated. Datcu and Rothkrantz [130] proposed an MSA method using facial and speech features. This method used HHM combined with LBP as the visual feature and MFCC as the speech feature for sentiment classification.

In late fusion methods, You et al. [131] combined vision and text to conduct MSA. They fine-tuned CNN to obtain visual features and used an unsupervised language model to learn the distributed representation of documents and paragraphs for MSA. For the problem of redundant information, Jiang et al. [132] proposed a fine-grained attention mechanism to interactively learn the cross-modal fusion representation of visual and text information. Due to the lack of systematically studied about the matching degree between cross-modal features at the emotional semantic level, Chen et al. [133] proposed a multimodal adaptive method for joint SA based on image–text relevance.

In hybrid fusion methods, Zhu et al. [134] learned the corresponding relationship between regions and words from the text–image pairs, introduced a cross-modal alignment module based on the cross-modal attention mechanism, and utilized an adaptive cross-modal gating module to fuse the multimodal features. To tackle three issues: 1) ignoring the object-level semantics in images; 2) primarily focusing on aspect–text and aspect–image interactions; and 3) failing to consider the semantic gap between text and image representations, Yu et al. [135] designed a general hierarchical interactive multimodal transformer (HIMT) model for aspect-based MSA.

D. Trimodal Sentiment Analysis

Trimodal SA aims to merge data from three modalities, text, video, and audio, to accurately predict sentiment. In this section, we categorize trimodal SA into multimodal representation learning and data fusion methods.

1) Multimodal Representation Learning-Based Methods: Pham et al. [136] proposed an MSA approach that used the seq2seq model, which performed unsupervised learning on the joint multimodal representation. The joint representation method requires all modes as input for representation learning, which is sensitive to noise or missing modes. Therefore, Pham et al. [137] explored a new method of joint representation and proposed a multimodal cyclic transformation network (MCTN), which learned robust joint multimodal representation

by the transformation between modes. Because MSA is based on unified multimodal annotation, existing methods are limited in capturing different sentiments across modalities. Yu et al. [138] proposed a label generation module based on self-supervised learning.

2) Multimodal Data Fusion-Based Methods: In early fusion methods, Castellano et al. [139] proposed a multimodal method that integrated the information from facial expressions, physical activity, gestures, and other information at the feature level for sentiment classification. Pérez-Rosas et al. [140] utilized word bags, OpenEAR, and CERT to fuse the features of text, acoustic, and visual modalities for MSA. Although the supplement of visual and speech modal information improves the classification accuracy, the related research [141] found that the text modality significantly impacts the classification results. Poria et al. [142] proposed a multikernel learning (SPF-GMKL) method to extract features from the text, which realized the detection of sentiment polarity from short video clips. On the basis of this work, Poria et al. [143] further discussed the role of the general framework for MSA and proposed a convolution neural network that used multiple kernel learning (MKL) for multimodal emotion recognition and analysis. Given three challenges of MSA in online opinion videos, Zadeh et al. [144] constructed a multimodal opinion-level sentiment intensity (MOSI) dataset that can be used for sentiment, subjectivity, and multimodal language research. Previous work is usually based on the assumption that the utterances in the video are independent of each other, ignoring the important role of context in identifying the sentiment of utterance. Therefore, Poria et al. [145] proposed an attention-based LSTM model for MSA. To combine cues from different modalities, Chen et al. [146] proposed a gated multimodal embedding LSTM [GME-LSTM(A)] with time attention.

In late fusion methods, Wöllmer et al. [147] proposed a decision-level fusion method for analyzing the sentiment of speakers in online videos. To solve the problem of modal conflict and redundant information, Majumder et al. [148] proposed a hierarchical method, which integrated the utterance feature vectors of different modal combinations. Each view from multimodal data has its own representation space and dynamics and contains some knowledge that other views cannot access. Therefore, to comprehensively and accurately describe multimodal data, Zadeh et al. [149] proposed a memory fusion network (MFN) for multiview sequential learning, which the delta memory attention network (DMAN) was designed to predict sentiment by fusing specific and cross-view information. To capture the contribution of different modalities for MSA, Akhtar et al. [150] proposed a contextual intermodal attention framework based on RNN, which used multimodal and contextual information to simultaneously predict the sentiment and emotion of discourse in multitask learning. MSA needs to take all modalities as inputs, and there will be modalities missing in the process of fusion. Tang et al. [151] proposed a coupled-translation fusion network (CTFN) that modeled bidirectional interaction through coupled learning to ensure robustness to missing modalities. Previous research on MSA focused on modal fusion and interaction, and had a lack of using the independence and correlation between modalities for dynamic MSA. Han et al. [152] proposed a bibimodal modality fusion for correlation-controlled MSA. As the classification ability of each modality is suppressed by single-task learning, Yang et al. [153] proposed a multimodal framework named two-phase multitask SA (TPMSA).

In hybrid fusion methods, Zadeh et al. [154] proposed a multiattention recurrent network (MARN), which used time clues to enhance the robustness of sentiment prediction. Although MARN used multiattention blocks to take advantage of the temporal interaction between modalities, this method was completely dependent on the attention mechanism so that it was very difficult to optimize the hyperparameters of its merged architecture. Verma et al. [155] proposed a deep higher order sequence fusion for MSA and performed multimodal fusion by extracting two kinds of contrast information from multimodal time series. Similar to the research of literature [150], Wang et al. [156] believed that not all modalities play the same role in SA and proposed an end2end fusion method with a transformer for MSA. Due to the heterogeneity of signals leading to the difference in distribution patterns, Hazarika et al. [157] proposed a modality-invariant and modality-specific representation for MSA, which learned the decomposition subspace of each mode and provided better representation as the input of fusion.

E. Model-Based Multimodal Data Fusion Methods

Traditional multimodal fusion can be divided into early fusion, late fusion, and hybrid fusion according to the fusion stages. With the development of deep learning technology, more and more models apply neural networks to different fusion stages for feature extraction and fusion. Therefore, multimodal fusion research can also be divided into model-agnostic and model-based fusion methods. Model-agnostic fusion means that the algorithm framework of multimodal fusion can be applied to any feature extraction and classification network. The process of modal fusion is independent of the specific model. The model-based fusion method is a model structure specially designed for specific tasks, such as visual question answering and multimodal dialog.

Zhu et al. [158] proposed a visual question-answering model with an attention mechanism. Although the attention mechanism allows attention to the visual content related to the problem, this simple mechanism is insufficient to model the complex reasoning features required for visual question answering or other high-level tasks. Therefore, Cadene et al. [159] proposed a multimodal relational reasoning model (MUREL) for visual question answering. The SA in the dialog system can help the system to understand the users' sentiments and produce a sympathetic response. However, the current work focuses on modeling the speaker and context information mainly on the text modality or only through feature connection to use multimodal information. In order to effectively carry out multimodal information fusion and capture long-distance context information, Hu et al. [160]

TABLE V
ANALYSIS AND COMPARISON OF DIFFERENT MULTIMODAL DATA FUSIONS

Structures	Advantages	Disadvantages	Models
Early fusion	Powerful ability to capture the interaction between different modalities.	Redundancy of data and time asynchronous.	[129], [130], [139], [140], [142], [143], [144], [145], [146]
Late fusion	Avoids the modal missing and overfitting problem.	Lacks of low-level interaction of multi-modal data, and more computationally intensive.	[131], [132], [133], [147], [148], [149], [150], [151], [152], [153]
Hybrid fusion	Combines the advantages of early and late fusion.	High computational complexity and difficult in training.	[134], [135], [154], [155], [156], [157]

proposed a new multimodal fusion graph convolution network (MMGCN) to establish edge connections between nodes corresponding to realize the interaction of context information.

F. Multimodal Alignment-Based Methods

In addition to the above methods, some studies have attempted to use multimodal alignment techniques for MSA. Multimodal alignment technology aims to establish a correspondence between different modalities that the information from different modalities can be aligned with each other [125].

Truong and Lauw [161] proposed a visual aspect attention network (VistaNet) that relied on visual information as alignment for pointing out the important sentences of a document using attention. To address the challenge of multimodal inherent data misalignment, Tsai et al. [162] introduced the Multimodal Transformer (MulT) to generically address the above issues in an end-to-end manner without explicitly aligning the data. As most existing methods mainly rely on combining the whole image and text while ignoring the implicit affective regions in the image, Li et al. [163] focused more on the alignment of multimodal fusion of visual and textual, and proposed a novel affective region recognition and fusion network for target-level multimodal sentiment classification.

G. Datasets and Performance Summary of Multimodal Sentiment Analysis

The construction of MSA datasets involves two steps: data collection and sentiment annotation. Data collection is generally selected from the network's movies, reviews, and videos. Most sentiment labels are manually labeled, and a small number is self-labeled. This section collects the datasets used for MSA in recent years according to the publication time and provides the performance summary of MSA, as shown in Tables A4 and A9 in the Supplementary Material.

VII. CHATBOTS AND CHATGPT IN SENTIMENT ANALYSIS

With the development of deep learning, researchers are leveraging SA techniques to empower chatbots with sentiment intelligence. Chatbot is a dialog system that interacts with humans via NLP technologies, and it aims to substitute human agents in answering questions, giving advice and providing sentiment support [164]. Ghosh et al. [165] proposed an LSTM to customize the degree of emotional content in generated sentences through an additional design parameter for generating conversational text. Hu et al. [166] designed a novel tone-aware chatbot that generated toned responses to user requests on social media. Adikari et al. [167] proposed an empathic conversational agent framework to detect and predict patient emotions to improve mental health and well-being outcomes.

Recently, a new AI-generated content (AIGC) product named Chat Generative Pre-trained Transformer (ChatGPT) has demonstrated amazing language understanding, generation, and knowledge reasoning capabilities [168]. ChatGPT is associated with chatbots, but it is not equivalent to them. In the field of SA, the main challenge lies in understanding semantics and context. Based on LLM, ChatGPT can leverage its language understanding and generation capabilities to assist SA. First, ChatGPT has good abilities of semantic understanding and knowledge reasoning, and it can help the model understand the contextual semantics and identify the sentiment reason to achieve a more accurate SA. Second, due to ChatGPT based on LLM, it can fine-tune the model on a large corpus of sentiment data to learn the patterns and nuances of sentiment expression. There are some LLM-based models that have been widely applied in SA, such as BERT, RoBERTa, and GPT-3.5. Xu et al. [169] introduced a review reading comprehension task and explored a novel posttraining approach on the popular language model BERT to enhance the performance of aspect-based SA tasks. Dai et al. [170] used the fine-tuned RoBERTa (FT-RoBERTa) to compare the induced trees from PTMs and the dependence parsing trees on several popular models for the ABSA task and showed that the FT-RoBERTa outperforms the best performance. Chen et al. [171] performed a comprehensive experimental analysis of GPT-3.5 covering nine popular natural language understanding (NLU) tasks that contained SA. However, ChatGPT also has some limitations in SA. For example, it fails to understand subtle and implicit sentiment expressions, such as sarcasm. Therefore, we need to combine multimodal information to improve the language understanding of ChatGPT for sentiment judgment.

VIII. CHALLENGES AND FUTURE TRENDS

In recent years, with the rapid development of deep learning, frameworks based on neural networks have been widely used in SA. However, there are different issues and challenges of SA tasks. In this section, we discuss open research challenges and provide three potential aspects to boost the SA tasks.

A. Open Research Challenges

SSA has been widely recognized and applied as the most extensive branch of SA. However, with the rapid development of internet technology, data are no longer limited to a single text modality but more in the form of multimodality. SSA fails to extract full information from multimodal data. As a new direction in SA, MSA develops rapidly and achieves better performance compare to SSA. At the same time, the network data have developed from a single to a diversified presentation, and MSA is more in line with the actual needs. In this section, we focus on the challenges of MSA.

1) Multimodal Representation: Good representation is a prerequisite for MSA. The challenges lie in the following: how to combine data from different sources, how to deal with different levels of noise, and how to deal with missing data.

2) Multimodal Fusion: The feature spaces of different modalities contain rich latent information, but how to fuse different modal information into a stable multimodal representation for downstream tasks is challenging. In addition, signals may not be aligned in time, and each modality shows different types and levels of noise.

3) Multimodal Alignment: Few datasets display dimension modal alignment. There may be one-to-one or one-to-many alignment types, and elements in one modality may not correspond to another modality.

4) Computational Resource: Training an accurate MSA model may require a large amount of computing power and storage resources; how to compress large models into smaller ones and achieve fast generalization with a small amount of data is a challenging task.

B. Future Trends

There are four challenges in MSA: multimodal representation, multimodal fusion, multimodal alignment, and computational resources. In this section, four potential aspects of these challenges are discussed to further improve MSA tasks.

1) Multimodal Pretraining Model: The first challenge of MSA is the construction of multimodal representation. The quality of multimodal representations directly determines the final performance of the model, and many methods attempt to construct multimodal representations. However, previous methods use separately pretrained visual and textual models that fail to capture the semantic and contextual relation of different modalities [127], [138]. The multimodal PTMs can learn the abundant knowledge of multimodal data to capture the semantic and contextual relation of different modalities and achieve good results with only a small amount of data through the learned parameters. Therefore, applying the PTM is one of the future research perspectives on MSA [172].

2) Framework for Maximum Mutual Information: The second challenge of MSA is multimodal data fusion. The previous studies attempt to fuse different modalities via early fusion, late fusion, and hybrid fusion [146], [152], [160]. However, these fusion methods lack control of the information flow from the original input to the fused embedding resulting in the loss of essential information and the introduction of the unexpected noise carried by individual modalities. As a possible solution, maximum mutual information can remove redundant information unrelated to downstream tasks and has excellent effects in capturing cross-domain information. Therefore, the introduction of maximum mutual information into multimodal data fusion and the realization of original information capture and noise removal are worthy of further research in MSA [173].

3) Cross-Modal Contrastive Learning: The third challenge of MSA is multimodal alignment. There are some studies that attempt to model multimodal alignment [161], [162], [163]. These methods usually utilize the attention mechanism to achieve multimodal alignment, but these methods fail to capture the interaction between different modalities resulting in poor modal alignment. The goal of CL is that all similar entities are in the adjacent regions of feature space, while all dissimilar entities are in the nonadjacent regions. Therefore, cross-modal CL is applied to MSA so that the distance between paired image text data and feature space is as close as possible, while the distance between nonpaired image text data and feature space is as far as possible. It is one of the future development directions in MSA to realize the semantic interaction and association of images and texts at different levels [174].

4) Knowledge Distillation and Few-Shot Learning: The fourth challenge of MSA is the computational resource. MSA requires a large number of computational resources to process a large amount of data and perform complex algorithmic calculations, and the optimization of MSA algorithms also requires a large number of computational resources and time. As a representative type of model compression and acceleration, knowledge distillation (KD) [175] compresses a large BERT model into a small student model while retaining the knowledge of the teacher model to reduce storage and computing costs and accelerate the reasoning processing. Meanwhile, FSL [176] is capable of learning from a very small number of samples and can generalize quickly by transforming and inducing limited information with prior knowledge to achieve rapid learning. Therefore, on the basis of compressing large models using KD, achieving fast generalization via annotation-efficient learning is a potential development direction.

IX. Conclusion

SA has attracted significant attention and application in the past decade, such as public opinion monitoring, esthetic analysis, and telephone service. This article provides a comprehensive survey of the current SA to provide guidance, reference, or potential insights and inspiration to researchers.

This work first takes the modal type as the thread to summarize and review the SA. We provide a novel framework to give researchers a more comprehensive understanding. Then, the workflow, trends, and datasets of SSA are introduced in detail, such as text, visual, and speech. Second, a new taxonomy is proposed to divide MSA into multimodal representation learning and data fusion, and multimodal fusion technologies and alignment methods are extended. Third, we discuss the

intelligent chatbots in SA. In addition, the open research challenges in different sentiment analyses are discussed. Finally, we introduce future directions, such as using cross-modal comparative learning to align multimodal data for MSA.

In addition, our research involves relatively little discussion on tasks that contain complex, implicit sentiment, such as multimodal sarcasm detection and multimodal fake news detection, which will be one of the future works.

REFERENCES

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.

[2] P. Ekman et al., "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Personality Social Psychol.*, vol. 53, no. 4, p. 712, 1987.

[3] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, Apr. 2013.

[4] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on visual sentiment analysis," *IET Image Process.*, vol. 14, no. 8, pp. 1440–1456, Jun. 2020.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[6] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, no. 1, pp. 3–14, Sep. 2017.

[7] T. Abdullah and A. Ahmet, "Deep learning in sentiment analysis: Recent architectures," *ACM Comput. Surveys*, vol. 55, no. 8, pp. 1–37, Aug. 2023.

[8] X. Han et al., "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, Jan. 2021.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[11] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2004, pp. 168–177.

[12] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An on-line lexical database," *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, 1990.

[13] S. Zhang, Z. Wei, Y. Wang, and T. Liao, "Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary," *Future Gener. Comput. Syst.*, vol. 81, pp. 395–403, Apr. 2018.

[14] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv:0205070*, 2002.

[15] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 412–418.

[16] M. Rezwanul, A. Ali, and A. Rahman, "Sentiment analysis on Twitter data using KNN and SVM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 1–7, 2017.

[17] J. Xu, D. Chen, X. Qiu, and X. Huang, "Cached long short-term memory neural networks for document-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1660–1669.

[18] Z.-Y. Dou, "Capturing user and product information for document level sentiment analysis with deep memory network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 521–526.

[19] Y. Zhang, J. Wang, and X. Zhang, "Conciseness is better: Recurrent attention LSTM model for document-level sentiment analysis," *Neurocomputing*, vol. 462, pp. 101–112, Oct. 2021.

[20] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 655–665.

[21] C. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proc. 25th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2014, pp. 69–78.

[22] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Exp. Syst. Appl.*, vol. 72, pp. 221–230, Apr. 2017.

[23] X. Wang, W. Jiang, and Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers*, 2016, pp. 2428–2437.

[24] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Dimensional sentiment analysis using a regional CNN-LSTM model," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 225–230.

[25] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Tree-structured regional CNN-LSTM model for dimensional sentiment analysis," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 28, pp. 581–591, 2020.

[26] D. Ma, S. Li, F. Wu, X. Xie, and H. Wang, "Exploring sequence-to-sequence learning in aspect term extraction," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3538–3547.

[27] Y. Yin, F. Wei, L. Dong, K. Xu, M. Zhang, and M. Zhou, "Unsupervised word and dependency path embeddings for aspect term extraction," 2016, *arXiv:1605.07843*.

[28] A. Giannakopoulos, C. Musat, A. Hossmann, and M. Baeriswyl, "Unsupervised aspect term extraction with B-LSTM and CRF using automatically labelled datasets," in *Proc. 8th Workshop Comput. Approaches to Subjectivity, Sentiment Social Media Anal.*, 2017, pp. 180–188.

[29] M. Venugopalan and D. Gupta, "An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis," *Knowl.-Based Syst.*, vol. 246, Jun. 2022, Art. no. 108668.

[30] B. Liu, T. Lin, and M. Li, "Boosting aspect category detection with inference heuristics and knowledge enhancement," *Knowledge-Based Syst.*, vol. 256, Nov. 2022, Art. no. 109855.

[31] K. Schouten, O. van der Weijde, F. Frasincar, and R. Dekker, "Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1263–1275, Apr. 2018.

[32] M. Hu et al., "Multi-label few-shot learning for aspect category detection," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6330–6340.

[33] H. Liu et al., "Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 1079–1087.

[34] Z. Fan, Z. Wu, X. Dai, S. Huang, and J. Chen, "Target-oriented opinion words extraction with target-fused neural sequence labeling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, 2019, pp. 2509–2518.

[35] A. P. B. Veyseh, N. Nouri, F. Dernoncourt, D. Dou, and T. H. Nguyen, "Introducing syntactic structures into target opinion word extraction with deep learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 8947–8956.

[36] S. Mensah, K. Sun, and N. Aletras, "An empirical study on leveraging position embeddings for target-oriented opinion words extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9174–9179.

[37] B. Huang and K. Carley, "Parameterized convolutional neural networks for aspect level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1091–1096.

[38] X. Wang, F. Li, Z. Zhang, G. Xu, J. Zhang, and X. Sun, "A unified position-aware convolutional neural network for aspect based sentiment analysis," *Neurocomputing*, vol. 450, pp. 91–103, Aug. 2021.

[39] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers*, 2016, pp. 3298–3307.

[40] Y. Liang, F. Meng, J. Zhang, J. Xu, Y. Chen, and J. Zhou, "A novel aspect-guided deep transition model for aspect based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5569–5580.

[41] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.

[42] S. Gu, L. Zhang, Y. Hou, and Y. Song, "A position-aware bidirectional attention network for aspect-level sentiment analysis," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 774–784.

[43] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent Twitter sentiment classification," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 49–54.

[44] T. H. Nguyen and K. Shirai, "PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2509–2514.

[45] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Recursive neural conditional random fields for aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 616–626.

[46] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 214–224.

[47] N. Majumder, S. Poria, A. Gelbukh, M. S. Akhtar, E. Cambria, and A. Ekbal, "IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3402–3411.

[48] N. Liu and B. Shen, "ReMemNN: A novel memory neural network for powerful interaction in aspect-based sentiment analysis," *Neurocomputing*, vol. 395, pp. 66–77, Jun. 2020.

[49] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107643.

[50] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4568–4578.

[51] R. Li, H. Chen, F. Feng, Z. Ma, X. Wang, and E. Hovy, "Dual graph convolutional networks for aspect-based sentiment analysis," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6319–6329.

[52] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proc. 22nd Nordic Conf. Comput. Linguistics*, 2019, pp. 187–196.

[53] Z. Wu and D. C. Ong, "Context-guided BERT for targeted aspect-based sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 16, 2021, pp. 14094–14102.

[54] X. Wang, M. Tang, T. Yang, and Z. Wang, "A novel network with multiple attention mechanisms for aspect-level sentiment analysis," *Knowledge-Based Syst.*, vol. 227, Sep. 2021, Art. no. 107196.

[55] Y. Zhou, L. Liao, Y. Gao, R. Wang, and H. Huang, "TopicBERT: A topic-enhanced neural language model fine-tuned for sentiment classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 380–393, Jan. 2023.

[56] A. Nazir and Y. Rao, "IAOTP: An interactive end-to-end solution for aspect-opinion term pairs extraction," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 1588–1598.

[57] S. Chen, J. Liu, Y. Wang, W. Zhang, and Z. Chi, "Synchronous double-channel recurrent network for aspect-opinion pair extraction," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6515–6524.

[58] L. Gao, Y. Wang, T. Liu, J. Wang, L. Zhang, and J. Liao, "Question-driven span labeling model for aspect–opinion pair extraction," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 14, May 2021, pp. 12875–12883.

[59] L. Xu, Y. K. Chia, and L. Bing, "Learning span-level interactions for aspect sentiment triplet extraction," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4755–4766.

[60] H. Peng, L. Xu, L. Bing, F. Huang, W. Lu, and L. Si, "Knowing what, how and why: A near complete solution for aspect-based sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 8600–8607.

[61] S. Chen, Y. Wang, J. Liu, and Y. Wang, "Bidirectional machine reading comprehension for aspect sentiment triplet extraction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 14, 2021, pp. 12666–12674.

[62] C. Zhang, L. Ren, F. Ma, J. Wang, W. Wu, and D. Song, "Structural bias for aspect sentiment triplet extraction," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 6736–6745.

[63] W. Zhang, Y. Deng, X. Li, Y. Yuan, L. Bing, and W. Lam, "Aspect sentiment quad prediction as paraphrase generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9209–9219.

[64] X. Bao, Z. Wang, X. Jiang, R. Xiao, and S. Li, "Aspect-based sentiment analysis with opinion tree generation," *Proc. in IJCAI*, 2022, pp. 4044–4050.

[65] T. Gao et al., "LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 7002–7012.

[66] Z. Chen and T. Qian, "Relation-aware collaborative learning for unified aspect-based sentiment analysis," in *Proc. 58th Annu. Meeting Assoc. for Comput. Linguistics*, 2020, pp. 3685–3694.

[67] X. Li, L. Bing, P. Li, and W. Lam, "A unified model for opinion target extraction and target sentiment prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 6714–6721.

[68] H. Luo, L. Ji, T. Li, D. Jiang, and N. Duan, "GRACE: Gradient harmonized and cascaded labeling for aspect-based sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2020, pp. 54–64.

[69] X. Zhang and X. Wang, "Novel survey on the color-image graying algorithm," in *Proc. IEEE Int. Conf. Comput. Inf. Technol. (CIT)*, Dec. 2016, pp. 750–753.

[70] J. Zhu, X. Wang, M. Chen, P. Wu, and M. J. Kim, "Integration of BIM and GIS: IFC geometry transformation to shapefile using enhanced open-source approach," *Autom. Construction*, vol. 106, Oct. 2019, Art. no. 102859.

[71] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[72] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 715–718.

[73] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 223–232.

[74] Y. Yang et al., "How do your friends on social media disclose your emotions?" in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, 2014, pp. 1–10.

[75] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–15.

[76] J. Wang, J. Fu, Y. Xu, and T. Mei, "Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks," in *Proc. IJCAI*, 2016, pp. 3484–3490.

[77] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7584–7592.

[78] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–17.

[79] J. Zhang, X. Liu, M. Chen, Q. Ye, and Z. Wang, "Image sentiment classification via multi-level sentiment region correlation analysis," *Neurocomputing*, vol. 469, pp. 221–233, Jan. 2022.

[80] Z. Shu-Ren, L. Xi-Ming, Z. Can, and Y. Qiu-Fen, "Facial expression recognition based on independent component analysis and hidden Markov model," *J. Image Graph.*, vol. 13, no. 12, pp. 2321–2328, 2008.

[81] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[82] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 6, pp. 636–642, Jun. 1996.

[83] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+3D facial action and expression recognition," *Pattern Recognit.*, vol. 43, no. 5, pp. 1763–1775, May 2010.

[84] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–187, Jan. 2007.

[85] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Trans. Affect. Comput.*, vol. 9, no. 3, pp. 343–350, Jul. 2018.

[86] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2401–2410.

[87] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.

[88] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[89] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 263–270.

[90] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.

[91] J. Cai et al., "Identity-free facial expression recognition using conditional generative adversarial network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1344–1348.

[92] Y. Shu, X. Gu, G.-Z. Yang, and B. Lo, "Revisiting self-supervised contrastive learning for facial expression recognition," in *Proc. BMVC*, 2022, pp. 1–14.

[93] D. Kim and B. C. Song, "Emotion-aware multi-view contrastive learning for facial emotion recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 178–195.

[94] Z. Zhao and Q. Liu, "Former-DFER: Dynamic facial expression recognition transformer," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1553–1561.

[95] H. Li, M. Sui, Z. Zhu, and F. Zhao, "NR-DFERNet: Noise-robust network for dynamic facial expression recognition," 2022, *arXiv:2206.04975*.

[96] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[97] Y. Xi, Q. Mao, and L. Zhou, "Weighted contrastive learning using pseudo labels for facial expression recognition," *Vis. Comput.*, pp. 1–12, Aug. 2022.

[98] C. Wang, J. Ding, H. Yan, and S. Shen, "A prototype-oriented contrastive adaption network for cross-domain facial expression recognition," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 4194–4210.

[99] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[100] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[101] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," in *Proc. Can. Conf. Electr. Comput. Eng.*, 1995, pp. 1062–1065.

[102] J. Morais, P. Bertelson, L. Cary, and J. Alegria, "Literacy training and speech segmentation," *Cognition*, vol. 24, nos. 1–2, pp. 45–64, Nov. 1986.

[103] P. Podder, T. Zaman Khan, M. Haque Khan, and M. Muktadir Rahman, "Comparative performance analysis of hamming, Hanning and blackman window," *Int. J. Comput. Appl.*, vol. 96, no. 18, pp. 1–7, Jun. 2014.

[104] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7378–7382.

[105] K. Garg and G. Jain, "A comparative study of noise reduction techniques for automatic speech recognition systems," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2016, pp. 2098–2103.

[106] M. Feraru and M. Zbancioc, "Speech emotion recognition for SROL database using weighted KNN algorithm," in *Proc. Int. Conf. Electron., Comput. Artif. Intell. (ECAI)*, Jun. 2013, pp. 1–4.

[107] S. A. Rieger, R. Muraleedharan, and R. P. Ramachandran, "Speech based emotion recognition using spectral feature extraction and an ensemble of KNN classifiers," in *Proc. 9th Int. Symp. Chin. Spoken Lang. Process.*, Sep. 2014, pp. 589–593.

[108] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[109] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jul. 2003, p. 401.

[110] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2007, pp. 1–21.

[111] H. K. Mishra and C. C. Sekhar, "Variational Gaussian mixture models for speech emotion recognition," in *Proc. 7th Int. Conf. Adv. Pattern Recognit.*, Feb. 2009, pp. 183–186.

[112] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *Proc. 5th Int. Conf. Knowl. Smart Technol. (KST)*, Jan. 2013, pp. 86–91.

[113] L. Sun, S. Fu, and F. Wang, "Decision tree SVM model with Fisher feature selection for speech emotion recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2019, no. 1, pp. 1–14, Dec. 2019.

[114] M. Jain et al., "Speech emotion recognition using support vector machine," 2020, *arXiv:2002.07590*.

[115] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. Interspeech*, 2015, pp. 1–4.

[116] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.

[117] G. Liu, W. He, and B. Jin, "Feature fusion of speech emotion recognition based on deep learning," in *Proc. Int. Conf. Netw. Infrastructure Digit. Content (IC-NIDC)*, Aug. 2018, pp. 193–197.

[118] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7390–7394.

[119] S. Kwon et al., "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Exp. Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114177.

[120] M. Li et al., "Contrastive unsupervised learning for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6329–6333.

[121] V. S. Alaparthi, T. R. Pasam, D. A. Inagandla, J. Prakash, and P. K. Singh, "ScSer: Supervised contrastive learning for speech emotion recognition using transformers," in *Proc. 15th Int. Conf. Human Syst. Interact. (HSI)*, Jul. 2022, pp. 1–7.

[122] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[123] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, Dec. 1966.

[124] D. A. Reynolds, "Gaussian mixture models," *Encyclopedia Biometrics*, vol. 741, nos. 659–663, pp. 1–5, 2009.

[125] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[126] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.

[127] G. Aguilar, V. Rozgic, W. Wang, and C. Wang, "Multimodal and multi-view models for emotion recognition," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 991–1002.

[128] Y. Zhang, Y. Zhang, W. Guo, X. Cai, and X. Yuan, "Learning disentangled representation for multimodal cross-domain sentiment analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 21, 2022, doi: 10.1109/TNNLS.2022.3147546.

[129] M. Wimmer, B. Schuller, D. Arsic, B. Radig, and G. Rigoll, "Low-level fusion of audio and video feature for multi-modal emotion recognition," in *Proc. 3rd Int. Conf. Comput. Vis. Theory Appl. VISAPP*, Funchal, Madeira, Portugal, 2008, pp. 145–151.

[130] D. Datcu and L. J. M. Rothkrantz, "Emotion recognition using bimodal data fusion," in *Proc. 12th Int. Conf. Comput. Syst. Technol. CompSysTech*, 2011, pp. 122–128.

[131] Q. You, J. Luo, H. Jin, and J. Yang, "Joint visual-textual sentiment analysis with deep neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1071–1074.

[132] T. Jiang, J. Wang, Z. Liu, and Y. Ling, "Fusion-extraction network for multimodal sentiment analysis," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2020, pp. 785–797.

[133] D. Chen, W. Su, P. Wu, and B. Hua, "Joint multimodal sentiment analysis based on information relevance," *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103193.

[134] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Trans. Multimedia*, early access, Mar. 16, 2022, doi: 10.1109/TMM.2022.3160060.

[135] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, early access, Apr. 28, 2022, doi: 10.1109/TAFFC.2022.3171091.

[136] H. Pham, T. Manzini, P. P. Liang, and B. Poczós, "Seq2Seq2Sentiment: Multimodal sequence to sequence models for sentiment analysis," in *Proc. Grand Challenge Workshop Human Multimodal Lang. (Challenge-HML)*, 2018, pp. 53–63.

[137] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 6892–6899.

[138] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, 2021, pp. 10790–10797.

[139] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*. Cham, Switzerland: Springer, 2008, pp. 92–103.

[140] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 973–982.

[141] E. Cambria, H. Wang, and B. White, "Guest editorial: Big social data analysis," *Knowl.-Based Syst.*, vol. 69, pp. 1–2, Oct. 2014.

[142] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2539–2544.

[143] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.

[144] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.

[145] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1033–1038.

[146] M. Chen, S. Wang, P. P. Liang, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 163–171.

[147] M. Wöllmer et al., "YouTube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, May 2013.

[148] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.

[149] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–8.

[150] M. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *Proc. Conf. North*, 2019, pp. 370–379.

[151] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5301–5311.

[152] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 6–15.

[153] B. Yang, L. Wu, J. Zhu, B. Shao, X. Lin, and T.-Y. Liu, "Multimodal sentiment analysis with two-phase multi-task learning," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 30, pp. 2015–2024, 2022.

[154] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–12.

[155] S. Verma et al., "Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 561–570.

[156] Z. Wang, Z. Wan, and X. Wan, "TransModality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, Apr. 2020, pp. 2514–2520.

[157] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131.

[158] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4995–5004.

[159] R. Cadene, H. Ben-younes, M. Cord, and N. Thome, "MUREL: Multimodal relational reasoning for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1989–1998.

[160] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5666–5675.

[161] Q.-T. Truong and H. W. Lauw, "VistaNet: Visual aspect attention network for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 305–312.

[162] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2019, p. 6558.

[163] L. Jia, T. Ma, H. Rong, and N. Al-Nabhan, "Affective region recognition and fusion network for target-level multimodal sentiment classification," *IEEE Trans. Emerg. Topics Comput.*, early access, Jan. 10, 2023, doi: 10.1109/TETC.2022.3231746.

[164] M. Firdaus, U. Jain, A. Ekbal, and P. Bhattacharyya, "SEPRG: Sentiment aware emotion controlled personalized response generation," in *Proc. 14th Int. Conf. Natural Lang. Gener.*, 2021, pp. 353–363.

[165] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, "Affect-LM: A neural language model for customizable affective text generation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 634–642.

[166] T. Hu et al., "Touch your heart: A tone-aware chatbot for customer care on social media," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2018, pp. 1–12.

[167] A. Adikari et al., "Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare," *Future Gener. Comput. Syst.*, vol. 126, pp. 318–329, Jan. 2022.

[168] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neur. Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.

[169] H. Xu, B. Liu, L. Shu, and S. Y. Philip, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 2324–2335.

[170] J. Dai, H. Yan, T. Sun, P. Liu, and X. Qiu, "Does syntax matter? A strong baseline for aspect-based sentiment analysis with RoBERTa," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2021, pp. 1816–1829.

[171] X. Chen et al., "How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks," 2023, *arXiv:2303.00293*.

[172] Y. Ling, J. Yu, and R. Xia, "Vision-language pre-training for multimodal aspect-based sentiment analysis," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2149–2159.

[173] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.

[174] Z. Lin et al., "Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 7124–7135.

[175] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[176] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–34, May 2021.

[177] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2004, pp. 271–278.

[178] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. ICML*, 2011, pp. 1–8.

[179] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, 2011, pp. 142–150.

[180] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 1014–1023.

[181] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.

[182] L.-C. Yu et al., "Building Chinese affective resources in valence-arousal dimensions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 540–545.

[183] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2014, pp. 27–35. [Online]. Available: https://aclanthology.org/S14-2004

[184] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 task 12: Aspect based sentiment analysis," in *Proc. 9th Int. Workshop Semantic Eval. (SemEval)*, 2015, pp. 486–495.

[185] M. Pontiki et al., "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proc. Int. Workshop Semantic Eval.*, 2016, pp. 19–30.

[186] J. S. Kessler, M. Eckert, L. Clark, and N. Nicolov, "The ICWSM 2010 JDPA sentiment corpus for the automotive domain," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media Data Workshop Challenge (ICWSM-DWC)*, 2010, pp. 1–10.

[187] L. Xu, H. Li, W. Lu, and L. Bing, "Position-aware tagging for aspect sentiment triplet extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 2339–2349.

[188] P. J. Lang et al., "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Center Study Emotion Attention, NIMH, Gainesville, FL, USA, Tech. Rep. A-8, 2005.

[189] B. Lu, M. Hui, and H. Yu-Xia, "The development of native Chinese affective picture system—A pretest in 46 college students," *Chin. Mental Health J.*, vol. 96, no. 11, pp. 719–722, 2005.

[190] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[191] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 83–92.

[192] E. S. Dan-Glauser and K. R. Scherer, "The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance," *Behav. Res. Methods*, vol. 43, no. 2, pp. 468–477, Jun. 2011.

[193] S. Jindal and S. Singh, "Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning," in *Proc. Int. Conf. Inf. Process. (ICIP)*. IEEE, 2015, pp. 447–451.

[194] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 526–540, Oct. 2018.

[195] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.

[196] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[197] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop Emotion Corpora Res. Emotion Affect (Satellite LREC)*. Paris, France, 2010, p. 65.

[198] J. M. Susskind, A. K. Anderson, and G. E. Hinton, "The Toronto face database," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. 3, 2010, p. 29.

[199] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2106–2112.

[200] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia-Mag.*, vol. 19, no. 3, pp. 34–41, Jul. 2012.

[201] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2013, pp. 117–124.

[202] J. Yang, M. Sun, and X. Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, Feb. 2017, pp. 1–11.

[203] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[204] A. Schmitt, S. Ultes, and W. Minker, "A parameterized and annotated spoken dialog corpus of the CMU lets go bus information system," in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*, 2012, pp. 3369–3373.

[205] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct. 2019.

[206] A. Adigwe, N. Tits, K. El Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," 2018, *arXiv:1806.09514*.

[207] J. James, L. Tian, and C. I. Watson, "An open source emotional speech corpus for human robot interaction applications," in *Proc. Interspeech*, 2018, pp. 2768–2772.

[208] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[209] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[210] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: An Italian emotional speech database," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 3501–3504.

[211] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "DEMoS: An Italian emotional speech corpus," *Lang. Resour. Eval.*, vol. 54, no. 2, pp. 341–383, Jun. 2020.

[212] N. Vrysas, R. Kotsakis, A. Liatsou, C. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *J. Audio Eng. Soc.*, vol. 66, no. 6, pp. 457–467, Jun. 2018.

[213] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "ShEMO: A large-scale validated database for Persian speech emotion detection," *Lang. Resour. Eval.*, vol. 53, no. 1, pp. 1–16, Mar. 2019.

[214] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, vol. 5, 2005, pp. 1517–1520.

[215] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech Lang., Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.

[216] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, 2018, pp. 88–93.

[217] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "The Mexican emotional speech database (MESD): Elaboration and assessment based on machine learning," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1644–1647.

[218] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[219] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, Nov. 2011, pp. 169–176.

[220] J. G. Ellis, B. Jou, and S.-F. Chang, "Why we watch the news: A dataset for exploring sentiment in broadcast video news," in *Proc. 16th Int. Conf. Multimodal Interact.*, 2014, pp. 104–111.

[221] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.

[222] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[223] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–5.

[224] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4619–4629.

[225] W. Yu et al., "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.

[226] L. Wang, W. Guo, X. Yao, Y. Zhang, and J. Yang, "Multimodal event-aware network for sentiment analysis in tourism," *IEEE MultimediaMag.*, vol. 28, no. 2, pp. 49–58, Apr. 2021.

[227] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2506–2515.

Qiang Lu (Graduate Student Member, IEEE) received the B.S. degree in computer science and technology and the M.S. degree from Shandong Jiaotong University, Jinan, China, in 2016 and 2021, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with Northwest University, Xi'an, China.

His research interests include natural language processing and sentiment analysis.

Xia Sun received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2006.

She is currently a Professor with the School of Information Science and Technology, Northwest University, Xi'an. She has coauthored 40 articles and is an editor or a coeditor of four books. Her current research interests include natural language processing and intelligent education.

Prof. Sun has reviewed many journals, including IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, *Pattern Recognition*, and *Chinese Journal of Electronics*.

Yunfei Long (Member, IEEE) received the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2019.

From 2019 to March 2020, he was a Research Fellow with the School of Medical Science, University of Nottingham, Nottingham, U.K. Since April 2020, he has been a Lecturer (Assistant Professor) with the School of Computer Science, University of Essex, Colchester, U.K. His research interests include innovative natural language processing (NLP) techniques; explainable artificial intelligence (AI) models for analyzing healthcare data, the utilization of lexical semantics in affective analysis, and other NLP applications; and the applications of NLP in media, legal, and crime studies.

Zhizezhang Gao received the B.S. and M.S. degrees in statistics from Beijing Normal University, Beijing, China, in 2019 and 2021, respectively. He is currently pursuing the Ph.D. degree with Northwest University, Xi'an, China.

His current research interests include education-related natural language processing (NLP), educational data mining, and medical statistics.

Jun Feng is currently a Professor with the School of Information Science and Technology, Northwest University, Xi'an, China. She is a coauthor of more than 100 articles and a coeditor of three books. Her recent projects have included medical image analysis with deep learning, intelligence education based on artificial intelligence (AI), and brain–human interaction. Her research areas include pattern recognition and machine learning, as well as their applications in different fields.

Tao Sun graduated in computer science and technology from the China University of Petroleum, Qingdao, China.

His current research interests include intelligence education.