

SURVEY

A Survey on Multimodal Aspect-Based Sentiment Analysis

HUA ZHAO¹, **MANYU YANG¹**, **XUEYANG BAI¹**, AND **HAN LIU**

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding author: Hua Zhao (huamolin@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0119501; in part by the National Natural Science Foundation of China under Grant 52374221; in part by the Science and Technology Development Fund of Shandong Province of China under Grant ZR2021MG038, Grant ZR2021QG038, Grant ZR2022MF288, and Grant ZR2023MF097; in part by the Taishan Scholar Program of Shandong Province under Grant ts20190936; and in part by the Shandong University of Science and Technology (SDUST) Intelligent Science and Security Governance Innovation Team.

ABSTRACT Multimodal Aspect-Based Sentiment Analysis (MABSA), as an emerging task in the field of sentiment analysis, has recently received widespread attention. Its aim is to combine relevant multimodal data to determine the sentiment polarity of a given aspect in text. Researchers have surveyed both aspect-based sentiment analysis and multimodal sentiment analysis, but, to the best of our knowledge, there is no survey on MABSA. Therefore, in order to assist related researchers to know MABSA better, we surveyed the research work on MABSA in recent years. Firstly, the relevant concepts of MABSA were introduced. Secondly, the existing research methods for the two subtasks of MABSA research (that is, multimodal aspect sentiment classification and aspect sentiment pairs extraction) were summarized and analyzed, and the advantages and disadvantages of each type of method were analyzed. Once again, the commonly used evaluation corpus and indicators for MABSA were summarized, and the evaluation results of existing research methods on the corpus were also compared. Finally, the possible research trends for MABSA were envisioned.

INDEX TERMS Multimodal aspect-based sentiment analysis, multimodal aspect sentiment classification, aspect sentiment pairs extraction.

I. INTRODUCTION

Sentiment analysis, also known as sentiment orientation analysis, opinion mining, etc., refers to the process of using natural language processing, machine learning, and deep learning related methods to process and analyze various modalities of data with sentiment orientations, such as text, image, and speech, in order to identify their sentiment tendencies. It has been one of the hot topics in the fields of natural language processing and image and video mining in recent years [1]. Sentiment analysis is widely used in fields such as e-commerce, online public opinion analysis, and intelligent customer service, and plays an important role in many cases. For individuals, it can help us better understand the underlying motivations behind human behavior and attitudes. For

enterprises, it can assist them in understanding user satisfaction and needs for products or services, thereby guiding decision-making and improving user experience. For the government, it can play an important role in understanding the public's response and emotions towards policies, in order to improve governance and public services.

Traditional sentiment analysis is mainly based on one of the modalities such as text [2], [3], [4], [5], image [6], [7], [8], and speech [9], [10]. However, when expressing sentiments, humans may synthetically adopt multiple forms such as texts, facial expressions, voices, and body languages to express sentiments. Therefore, there are certain limitations to depending on single modality for sentiment analysis. Using multimodal data synthesis to determine sentiments is one of the better solutions to this limitation, which has attracted more and more researchers to strive for it. Specifically, various methods such as modality representation [11], [12], modality alignment [13], [14], and modality fusion [15], [16], [17] are

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

explored to effectively improve the above limitations in single modality sentiment analysis.

Aspect-based sentiment analysis is a subtask in sentiment analysis, which is a fine-grained task aimed at analyzing the sentiment polarity of different aspects (i.e., entities) in the text. In recent years, a large amount of research has been invested in aspect-based sentiment analysis [18], [19], [20], [21], [22], [23], and good results have been achieved. MABSA is based on traditional aspect-based sentiment analysis, integrating information from multiple modalities for sentiment analysis. Its goal is to combine relevant multimodal data to determine the sentiment polarity of a given aspect in the text. MABSA mainly includes three subtasks, namely multimodal aspect term extraction, multimodal aspect sentiment classification, and aspect sentiment pairs extraction.

Some scholars have surveyed the existing work on aspect-based sentiment analysis [24], [25], [26], [27], [28] and multimodal sentiment analysis [29], [30], [31], [32]. However, to our knowledge, the existing work on MABSA has not yet been sorted and summarized. Therefore, this paper focuses on summarizing the existing research methods for MABSA, and analyzing the advantages and disadvantages of the existing methods. In addition, the commonly used evaluation corpus and indicators for MABSA tasks, as well as the evaluation results of existing methods on the corpus, are also summarized. Finally, the possible research trends for MABSA are envisioned.

II. OVERVIEW OF RELATED CONCEPTS

A. SENTIMENT ANALYSIS

Sentiment Analysis (SA), also known as opinion mining, orientation analysis, etc., refers to the usage of computer technology to analyze data such as text, speech, or image to infer the sentiments or emotional states expressed therein [33]. It is an important research task in natural language processing, aiming to enable computers to understand and obtain human sentiment.

Sentiment analysis includes sentiment classification, emotion analysis, opinion extraction, comment mining, etc. Among them, sentiment classification is the most widely studied issue. According to different granularity, sentiment classification can be divided into document level, sentence level, and aspect level [34]. Document level sentiment classification aims to predict the overall sentiment polarity of an entire document or a longer segment of text. Sentence level sentiment classification predicts sentiment polarity for each sentence. Compared to document level or sentence level sentiment classification, aspect level sentiment classification focuses on predicting the sentiment polarity of specific target aspects. It analyzes sentiments related to specific aspects, goals, or themes in the text. See section B for more detailed introduction about it.

B. ASPECT-BASED SENTIMENT ANALYSIS

Aspect-Based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis task, which aims at extracting opinions

about specific aspects from a large amount of unstructured text. It is usually classified into three sentiment polarities: positive, negative, and neutral.

For example, the text *‘The food is great but the service and the environment are dreadful’* contains three aspects, namely *‘food’*, *‘service’* and *‘environment’*, as well as two opinion words *‘great’* and *‘dreadful’*, corresponding to positive, negative, and negative sentiment polarities. Through fine-grained aspect-based sentiment analysis, we can accurately determine the sentiment polarity of each aspect in specific contexts, thereby obtaining deeper and more detailed sentiment understanding. This is crucial for us to comprehensively evaluate and understand the subtle differences in sentiment expression, as well as people’s attitudes and sentiment tendencies towards specific aspects.

C. MULTIMODAL SENTIMENT ANALYSIS

Multimodal Sentiment Analysis (MSA) is a task that utilizes multiple modal data for sentiment analysis, such as text, image, etc. Compared with single mode sentiment analysis, multimodal sentiment analysis can obtain more comprehensive and accurate sentiment information from different perception channels. For example, the sentence *“Today’s weather is really nice!”* expresses positive sentiment when analyzed solely from the text, but when combined with an image of a rainy day, the overall sentiment is negative sentiment with ironic connotations. For this situation, it is difficult to determine the sentiment solely based on the text modality [35]. Certainly, in some cases, the inclusion of image modality in multimodal sentiment analysis can introduce certain amount of noise. For instance, when the text contains words like *‘sad’* or *‘unhappy’* while the associated image shows a smiling face expressing positive sentiment, the overall sentiment may actually be negative. In such cases, the image modality adds a certain level of noise that can impact the accurate determination of the overall sentiment.

Although multimodal data contains richer information, how to effectively integrate multimodal information is a key issue in current multimodal sentiment analysis tasks. The modal fusion methods is generally divided into three types, which are feature layer fusion, decision layer fusion, and hybrid fusion [36], and the basic ideas of them is shown in Figure 1. Feature layer fusion refers to the direct concatenation or weighted connection of feature vectors of different modalities into a new vector, which is then input into the classifier for sentiment analysis, as shown in Figure 1 (a). Decision layer fusion refers to the independent classification of features from different modalities, followed by weighting, voting mechanism, and other processing to generate the final decision, as shown in Figure 1 (b). Hybrid fusion is the integration of feature layer fusion and decision layer fusion methods. As shown in Figure 1 (c), a classification result is firstly obtained using feature layer fusion, and then it is fused with the result of another classifier using decision layer fusion. These three fusion methods each have their own

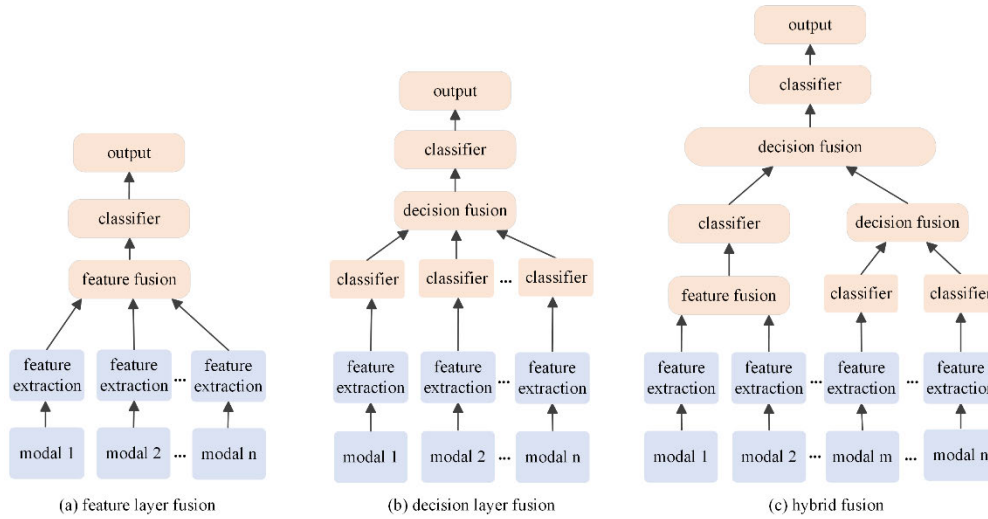


FIGURE 1. Schematic diagram of three multimodal fusion methods.


Image		
Text	a. Donald Trump rejects entire intelligence community in new defense of Russia	b. This is what Harry Potter 's grown – up family looks like
Output	(Donald Trump , negative) (Russia , neutral)	(Harry Potter , positive)

FIGURE 2. Example of MABSA task.

strengths, and it is necessary to choose the appropriate fusion method based on the actual task requirements.

D. MULTIMODAL ASPECT-BASED SENTIMENT ANALYSIS

Multimodal aspect-based sentiment analysis is a task that also utilizes multiple modal data for sentiment analysis, but it is oriented towards finer grained sentiment analysis. Its goal is to extract aspects and corresponding sentiment tendencies in the text by combining data from multiple modalities.

At present, MABSA related research mainly integrates two different modalities, which are text and image. Taking an example from Twitter-17 dataset, as shown in Figure 2 (a), we expect to extract two aspect-sentiment pairs from text-image pair, namely (**Donald Trump**, negative) and (**Russia**, neutral). For **Donald Trump**, it can be seen from the text that the sentiment expressed is negative, while the image associated with it expresses positive sentiment, which may bring some noise. In Figure 2 (b), we want to extract the (**Harry**

Potter, positive) pair. It is difficult to determine the sentiment polarity of the **Harry Potter** solely through textual information, but the image associated with it expresses positive sentiment, which provides important clues. As an emerging sentiment analysis subtask, how to effectively achieve cross-modal alignment between image and text, and solve the inconsistency between image and text, still faces serious challenges.

III. EXISTING METHODS FOR MABSA

MABSA includes three main subtasks: Multimodal Aspect Terms Extraction (MATE), Multimodal Aspect Sentiment Classification (MASC), and Aspect Sentiment Pairs Extraction (ASPE), which is a combination of MATE and MASC. At present, research on MABSA mainly focuses on MASC and ASPE. Therefore, this section focuses on summarizing the relevant methods of MASC and ASPE, and Figure 3 gives the timelines of them.

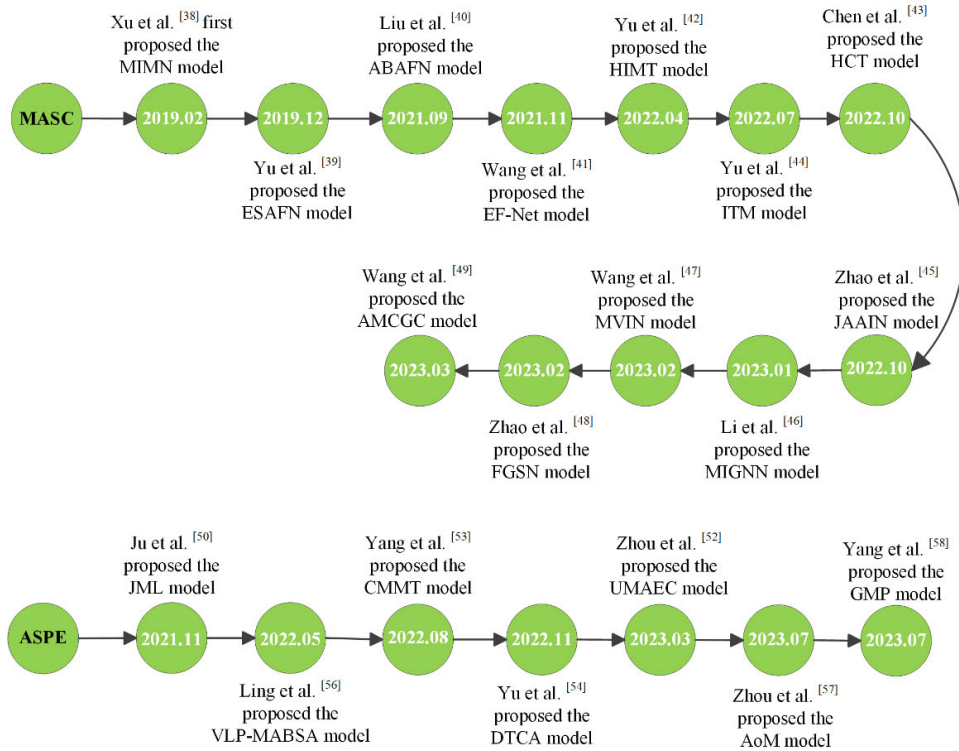


FIGURE 3. Development timeline of related research.

A. MULTIMODAL ASPECT SENTIMENT CLASSIFICATION METHODS

Given a text S containing n words $S = \{w_1, w_2, \dots, w_n\}$, l associated images $I = \{V_1, V_2, \dots, V_l\}$, and m aspects $A = \{A_1, A_2, \dots, A_m\}$, where w_i represents the i^{th} word, V_i represents the i^{th} image, A_i represents the i^{th} aspect in the text. Taking the pair (S, I) and one of the aspect A_i as input, the goal of MASC is to learn a sentiment classifier that maps (S, I, A_i) to \mathcal{Y} , i.e. $f(S, I, A_i) \rightarrow \mathcal{Y}$, where $\mathcal{Y} \in \{positive, negative, neutral\}$ or an integer sentiment score set from 1 to 10, which is determined based on different datasets.

According to existing references, the existing MASC methods can be roughly divided into two categories, namely attention mechanism-based MASC method and graph convolutional network-based MASC method.

1) ATTENTION MECHANISM-BASED MASC METHOD

Attention mechanism is a model used to simulate human attention behavior, which was first introduced in computer vision in 2015 to enhance key information extraction in image or video processing [37]. Its main idea is to focus on information related to the current task and ignore irrelevant information when processing information.

At present, attention mechanism has been widely applied in many fields such as computer vision and natural language processing. In MASC research, the use of attention mechanism can help models focus on aspect related information, thereby extracting the most relevant content from text or

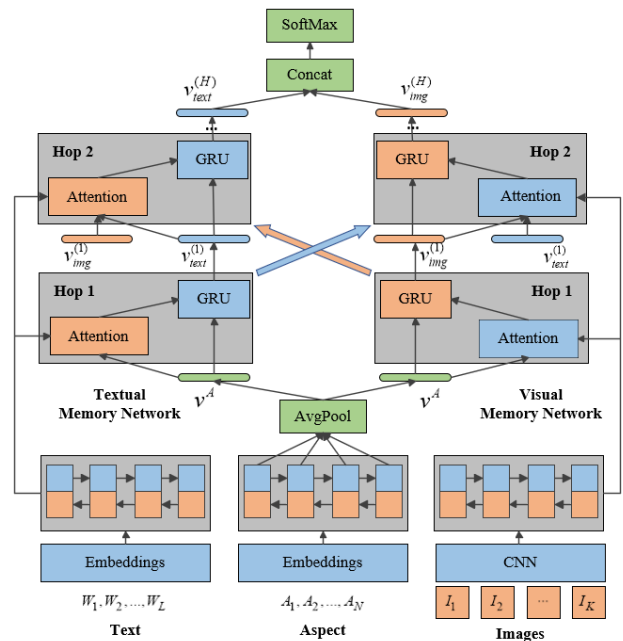


FIGURE 4. MIMN model proposed by Xu et al. [38].

image. For text, attention mechanism can identify the words or phrases most relevant to the aspect, ignoring content unrelated to the aspect. For image, attention mechanism can assign attention weights to aspect related image regions to capture

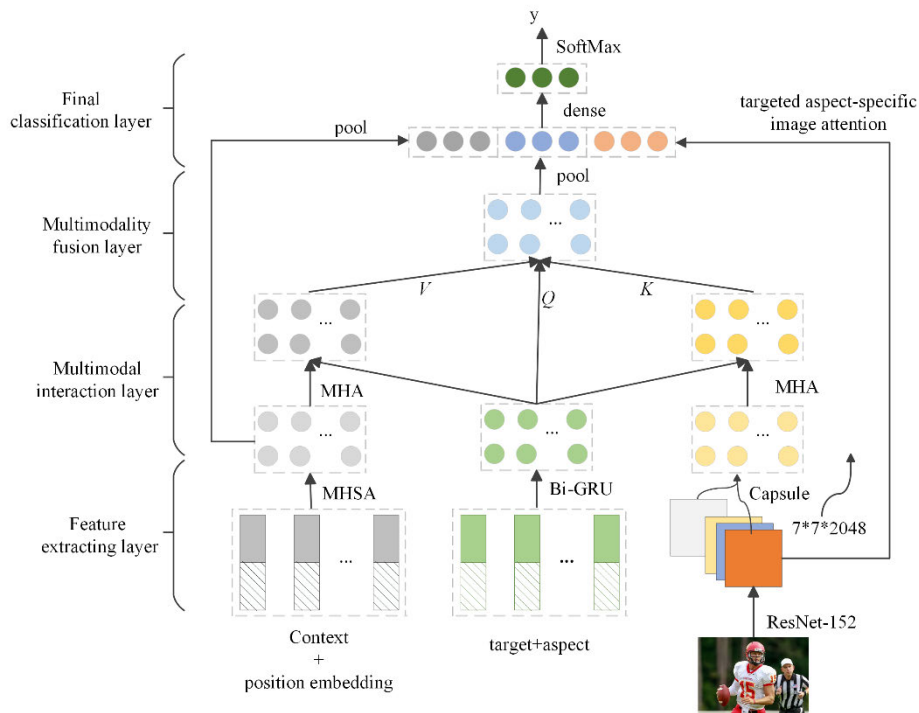


FIGURE 5. EF-Net model proposed by Wang et al. [41].

visual features related to sentiment classification. In addition, in cross-modal interaction between text and image, the use of attention mechanism can help the model select and align important information between text and image, thereby better capturing and understanding aspect related content in cross-modal data.

Xu et al. [38] proposed a model called Multi-Interactive Memory Network (MIMN), whose basic framework is shown in Figure 4. The MIMN model includes two interactive memory networks to monitor the textual and visual information of a given aspect, thereby learning not only the interactive effects between cross-modal data, but also the self-effects in single modal data. Specifically, after obtaining the features of the aspect, text and images, the aspect-guided attention mechanism is adopted to obtain text and image representations with aspect information. Then, multiple interactive attention modules are used to obtain interactive representations between the two modalities, and Gate Recurrent Unit (GRU) is used to update new textual and visual features for the next step of operation. Finally, the final output of GRU is used as the final text and visual features, and concatenated into the SoftMax layer for prediction.

Yu et al. [39] proposed the Entity Sensitive Attention and Fusion Network (ESAFN) model for MABSA tasks. ESAFN adds $\langle e \rangle$ and $\langle /e \rangle$ flags before and after the target entity, dividing the input text into three parts: left context, right context, and target entity, and using attention mechanism to generate entity sensitive text representations for each left and right context. In the text fusion layer, a low rank bilinear

pooling operator is used to model the interaction between entities and contexts (both left and right), and original context is added as the final text feature. In addition, the ESAFN model learns entity sensitive visual representations through entity oriented visual attention mechanism, and filters visual noise through gating mechanism. Finally, in the multimodal fusion layer, another bilinear pooling operator is used to capture the interaction between text and visual modalities, and both text feature presentation and visual feature representation are introduced as the final multimodal representation input to the SoftMax layer for prediction.

Liu et al. [40] proposed a model called Aspect-Based Attention and Fusion Network (ABAFN). This model utilizes attention mechanism to weight contextual and visual representations based on aspects, and then cascades and fuses the weighted representations of the two modalities to perform sentiment label classification tasks.

Gu et al. [41] designed an Attention Capsule Extraction and Multi-head Fusion Network (EF-Net) for MABSA, the basic framework of which is shown in Figure 5. EF-Net extracts image features using ResNet-152 and inputs them into a single layer capsule network to obtain the position information of the target in the image. Then, it uses a multi-head attention mechanism to obtain target specific textual attention and target specific visual attention, and uses a multi-head attention mechanism to achieve the fusion of multimodal features. Finally, the text features and multimodal features are averaged and cascaded with the target specific visual attention to obtain the final multimodal representation. The final

multimodal feature representation is linearly transformed and then fed into the SoftMax layer for sentiment classification.

Yu et al. [42] proposed a model called Hierarchical Interactive Multimodal Transformer (HIMT). This model proposes a hierarchical interaction module that first utilizes the aspect aware Transformer layer to obtain aspect aware text and image representations, and then models deep modal interactions using the multimodal fusion Transformer layer. This module also enhances the aspect aware text or image representations through self-attention mechanism. Chen et al. [43] proposed a Hierarchical Cross-modal Transformer (HCT) neural network model. This model designs a multimodal interaction module based on a cross-modal Transformer to model the interaction between text and image, thereby obtaining text related image representations and image related text representations. These two representations are then concatenated into a standard Transformer structure to model the interaction between these two representations, resulting in the final multimodal fusion representation and fed into the SoftMax layer for sentiment classification.

Yu et al. [44] proposed a coarse-to-fine grained Image-Target Matching (ITM) network. After extracting features, ITM firstly uses a coarse-grained matching module to capture the image-target relevance and alleviate the noise from unrelated image. Secondly, the fine-grained matching module further identifies the fine-grained visual objects aligned with the input target in those target-related image. Finally, the image-based target representation generated by the fine-grained matching module is cascaded with the text representation, fed into the Transformer layer for multimodal fusion, and then fed into the SoftMax layer for sentiment classification.

Zhao et al. [45] proposed an image, text and aspect sentiment recognition method based on a Joint Aspects Attention Interaction Network (JAAIN). This method addresses the inconsistency and correlation of image and text data. By multi-level fusion of aspect information and image and text information, image and text unrelated to a given aspect are removed, and the sentiment representation of the modal data of a given aspect is enhanced. The sentiment representations of text data, image data, and aspect sentiments are concatenated, fused, and fully connected to achieve sentiment discrimination.

2) GRAPH CONVOLUTIONAL NETWORK-BASED MASC METHOD

Graph Convolutional Network (GCN) is a deep learning model specifically designed for processing graph data. The core idea of GCN is to perform convolution operations on the graph structure, utilizing local domain information between nodes and their neighboring nodes for feature extraction. By performing convolution operations on connected nodes, GCN can encode local information in the graph, and through multi-layer GCN operations, each node can learn global information.

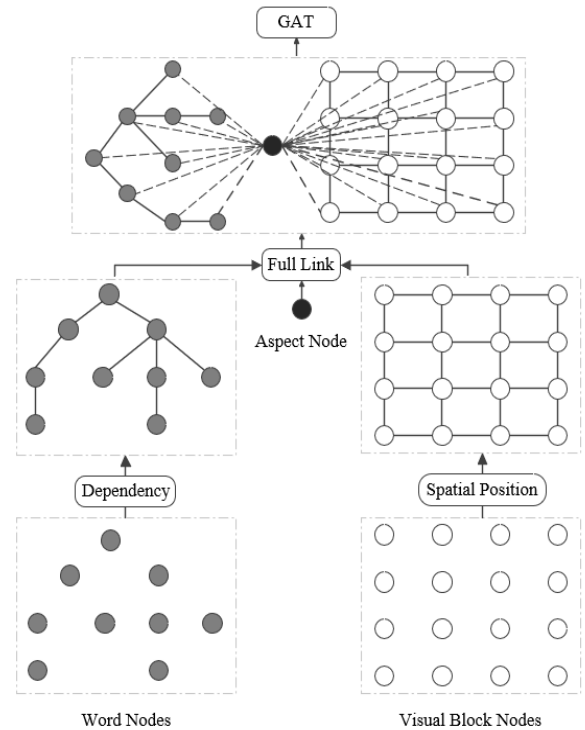


FIGURE 6. MIGNN model proposed by Li et al. [46].

Li and Li [46] proposed a Modal Interaction Graph Neural Network (MIGNN), whose basic framework is shown in Figure 6. This network connects semantic units of different modalities using aspect to form a multimodal interaction graph, and then utilizes the message passing mechanism in the Graph Attention Network (GAT) to fuse information from different data sources. The node in the multimodal interaction graph are fine-grained semantic units of each modal data, such as text words and visual blocks. Among them, the edges between text words are grammatical dependencies, the edges between visual blocks are spatial positional relationships, and aspect is fully connected to multimodal semantic units. Finally, the information from text and image is aggregated into the representation of aspect node through edges between each node, and the aspect node representation from the final layer of GAT output is input into the SoftMax layer for classification.

Wang et al. [47] proposed a Multiview Interaction Learning Network (MVIN) model, whose basic framework is shown in Figure 7. The MVIN model extracts features from both contextual and syntactic views of text, in order to fully utilize the global features of the text during multimodal interaction. Then model the relationship between text, image, and aspect to achieve multimodal interaction; Simultaneously integrating interactive representations of different modalities, dynamically obtaining the contribution of visual information to each word in the text, and fully extracting the correlation between modalities and aspect. Finally, the fused features are

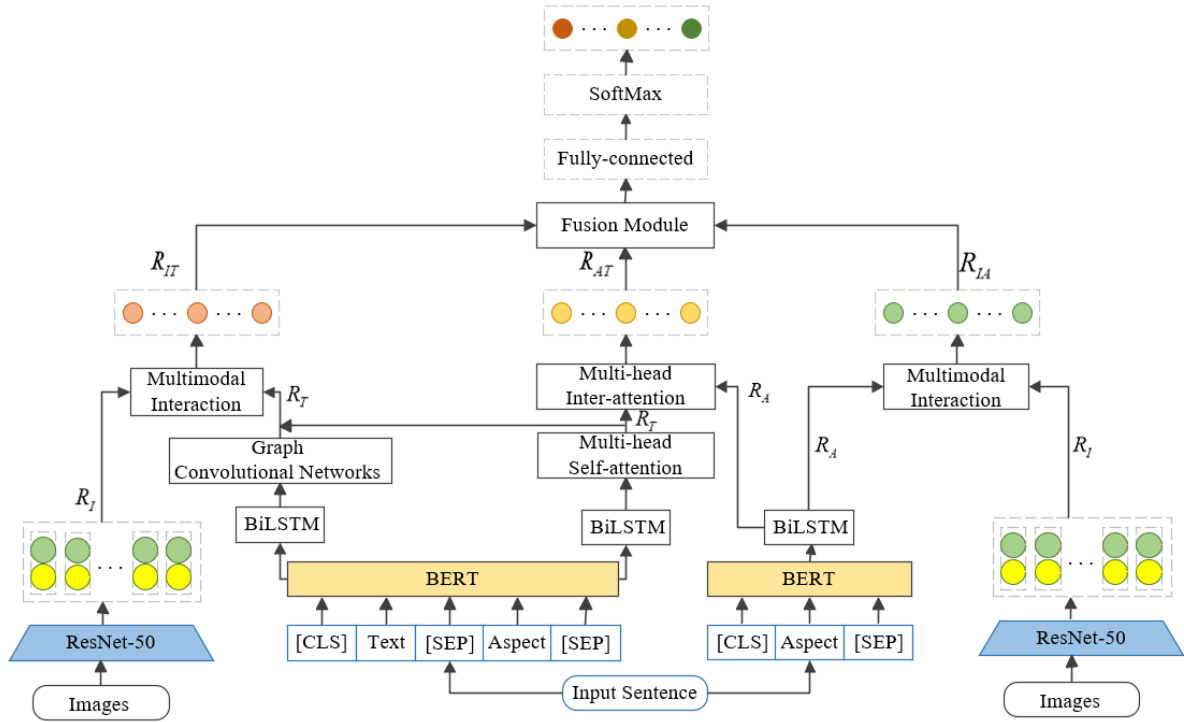


FIGURE 7. MVIN model proposed by Wang et al. [47].

input into the fully connected layer and SoftMax layer for sentiment classification.

Zhao and Yang [48] proposed a Fusion with GCN and SE-ResNeXt Network (FGSN) model. This model constructs a graph convolutional network on the dependency tree of a text, utilizes syntactic information and word dependencies to obtain contextual and aspect representations, and utilizes positional attention and channel attention mechanism to obtain image features. Then, image features and text features are fused for sentiment polarity classification. Wang et al. [49] proposed an Aspect-level Multimodal Co-attention Graph Convolutional (AMCGC) sentiment analysis model, which utilizes the self-attention mechanism of orthogonal constraints to generate semantic maps of each modality. Then, through graph convolution and bidirectional gated local cross-modal interaction mechanism, fine-grained cross-modal correlation and mutual alignment are gradually achieved.

3) COMPARISONS OF THE EXISTING MASC METHODS

Everything has two sides. The attention mechanism-based MASC method and the GCN-based MASC method have their own advantages and disadvantages in MASC, as follows:

Firstly, the attention mechanism-based MASC method can selectively focus on important information of different modalities and capture the interactions between different modalities, thereby better understanding and modeling the problems of multimodal sentiment analysis. In addition, the attention weight of the model can be used to explain the aspect

words or modalities that the model focuses on in classification tasks, which helps improve the interpretability and visualization ability of the model. However, the attention mechanism requires sufficient training data to learn effective attention weight distribution. If the dataset is small or unbalanced, the model may not be able to accurately learn the appropriate attention distribution, thereby affecting model performance.

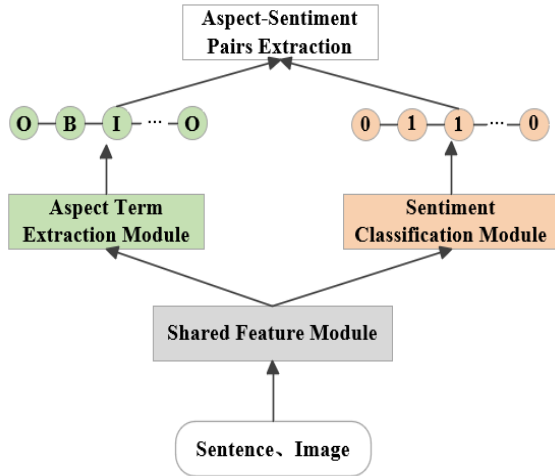
Secondly, the GCN-based MASC method can effectively utilize graph structure information to model nodes and edges in multimodal data, making them suitable for processing data with complex relationships and structures. In MABSA, there are rich correlations and dependencies between modalities, and GCN can better handle this complexity. However, GCN involves computing and storing the entire graph data, which may incur significant computational and storage overhead. GCN also has a certain dependence on the quality of the input graph structure. If the connection relationship of graph data is inaccurate or incomplete, it may affect the performance of the model.

B. ASPECT SENTIMENT PAIRS EXTRACTION METHODS

Given a text containing n words $S = \{w_1, w_2, \dots, w_n\}$ and related image V , the goal of aspect sentiment pairs extraction is to extract all aspects contained in the text and their corresponding sentiment categories, namely $\{a_1^s, a_1^e, s_1, \dots, a_i^s, a_i^e, s_i, \dots, a_m^s, a_m^e, s_m\}$, where a_i^s, a_i^e, s_i represents the starting position, ending position, and corresponding sentiment category of the i^{th} aspect, respectively, while m represents the number of aspects in the text.

TABLE 1. Example of joint-based ASPE method and unified-based ASPE method.

Text	Screen	clarity	are	great	,	but	battery	is	not	durable	.
Joint-based	B	E	O	O	O	O	S	O	O	O	O
	POS	POS	O	O	O	O	NEG	O	O	O	O
Unified-based	B-POS	E-POS	O	O	O	O	S-NEG	O	O	O	O

**FIGURE 8.** UMAEC model proposed by Zhou et al. [52].

According to existing references, there are four main research methods for ASPE, namely pipeline-based ASPE method, joint-based ASPE method, unified-based ASPE method, and text generation-based ASPE method [28].

1) PIPELINE-BASED ASPE METHOD

The pipeline-based ASPE method treats MATE and MASC as two independent subtasks, and implements the ASPE in a pipeline manner, that is, executing the MATE task first and then the MASC task. This pipeline approach is simple to implement, but inefficient and prone to error propagation.

Ju et al. [50] jointly performed MATE and MASC for the first time and proposed a Joint Multimodal Learning (JML) method to assist in cross-modal relationship detection. JML constructs an auxiliary text-image relation detection module to control the reasonable utilization of visual information, and then uses a layered framework to bridge the multimodal connection between MATE and MASC, and visually guides each sub-module separately. Finally, all aspects of sentiment polarity that rely on joint extraction are obtained.

2) JOINT-BASED ASPE METHOD

The joint-based ASPE method treats the MATE and MASC subtasks as two sequence labeling problems, and trains these two subtasks through a multi-task learning framework to utilize and interact with their relationship information. The final result is obtained by combining the predicted results of these two subtasks [51].

Zhou et al. [52] proposed a unified framework for MATE and MASC (UMAEC), and the overall framework of the model is shown in Figure 8. UMAEC first establishes a shared feature module to model potential semantic associations between tasks, and then adopts sequence annotation to simultaneously output multiple aspects and their corresponding sentiment categories contained in the text. Finally, a simple algorithm is used to implement ASPE.

3) UNIFIED-BASED ASPE METHOD

The unified-based ASPE method treats the MATE and MASC subtasks as sequence labeling problems based on a unified labeling scheme, ignoring the boundaries of these two subtasks and treating them as a sequence labeling task, using a unified labeling scheme such as B-POS [51]. Table 1 provides an example of joint-based ASPE method and unified-based ASPE method.

Yang et al. [53] proposed a multi-task learning framework named Cross-Modal Multitask Transformer (CMMT), which included two auxiliary tasks to learn intra-modal representations of aspects or sentiment perception, and introduces a text-guided cross-modal interaction module to dynamically control the contribution of visual information to each word representation in inter-modal interactions. The obtained multimodal representation is then fed to a standard CRF layer to predict the label sequence. Yu et al. [54] proposed a Dual-encoder Transformer with Cross-modal Alignment (DTCA), which introduces two auxiliary tasks to enhance cross attention performance and proposes minimizing the Wasserstein distance between the two modalities to align text and image, and feed the obtained multimodal feature to a standard CRF layer prediction label sequence.

4) TEXT GENERATION-BASED ASPE METHOD

The generative pre-training model Bidirectional and Auto-Regressive Transformers (BART) uses a standard Transformer based sequence to sequence structure, which combines a bidirectional Transformer encoder and a unidirectional autoregressive Transformer decoder to pre-train input text containing noise for denoising reconstruction. It is a typical denoising autoencoder [55]. In recent research, some scholars have successfully applied BART to MABSA tasks, transforming ASPE task into text generation task, and achieved good results.

Ling et al. [56] proposed a task specific Vision-Language Pre-training framework for MABSA (VLP-MABSA), which

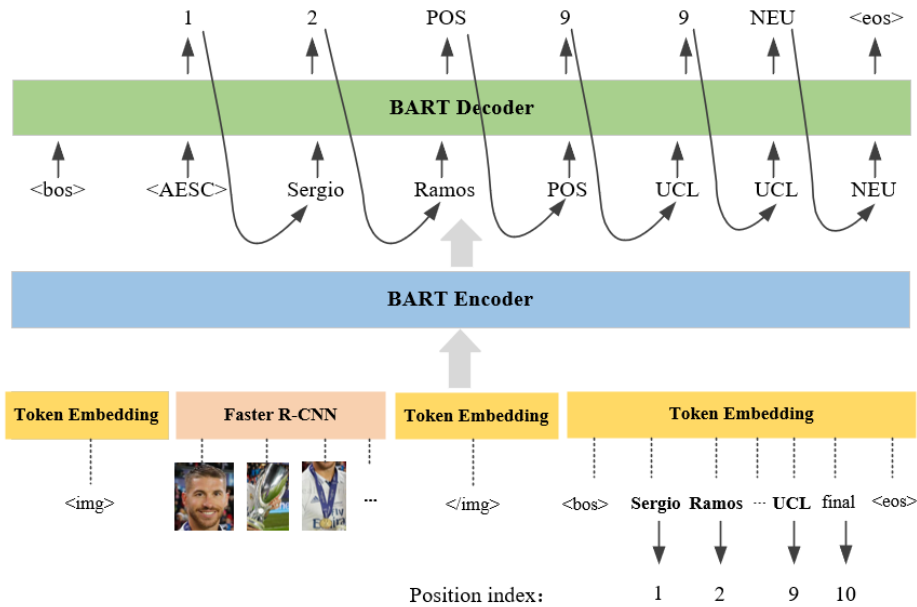


FIGURE 9. Example of VLP-MABSA model proposed by Ling et al. [56] in downstream ASPE task.

was a unified multimodal encoder-decoder architecture based on BART for all pre-training and downstream tasks. In addition to the general Masked Language Modeling (MLM) and Masked Region Modeling (MRM) tasks, three task-specific pre-training tasks were further introduced, including Textual Aspect-Opinion Extraction, Visual Aspect-Opinion Generation, and Multimodal Sentiment Prediction, to identify fine-grained aspect, opinions, and their cross-modal alignments. An example of this model for ASPE task is shown in Figure 9.

Zhou et al. [57] proposed an Aspect-oriented Method (AoM) to detect semantic and sentiment information related to aspects. This method designs an aspect-aware attention module between the BART based encoder-decoder architecture to simultaneously select text tags and image blocks related to aspect semantics, and explicitly introduces sentiment embedding into AoM. Then, graph convolutional network is used to model visual-text and text-text interaction. Yang et al. [58] first built diverse and comprehensive multimodal few-shot datasets according to the data distribution, and then proposed a novel Generative Multimodal Prompt (GMP) model for MABSA, which includes the Multimodal Encoder module and the N-Stream Decoders module. Furthermore, a subtask was introduced to predict the number of aspects in each instance to construct the multimodal prompt.

5) COMPARISONS OF THE EXISTING ASPE METHODS

The four ASPE methods mentioned above have their own advantages and disadvantages, which are as follows:

Firstly, the pipeline-based ASPE method is simple and straightforward, easy to implement and understand. However,

this method uses two completely independent models to implement ASPE step by step, ignoring the potential semantic associations between the two tasks. In addition, the MATE model extracts multiple aspects of the text at once, while the MASC model can only predict the sentiment polarity of one aspect at a time. The throughput of the former is greater than that of the latter, and MASC must be performed after the MATE is completed, resulting in low ASPE efficiency.

Secondly, the joint-based ASPE method can fully utilize the correlation and dependency relationship between two sub-tasks, improve the performance and generalization ability of the model, and also share common features for representation and learning. However, the training process of this method may be more complex, requiring the design of appropriate joint loss functions and training strategies.

Thirdly, the unified-based ASPE method can better capture the relationships and interactions between two subtasks and improve overall performance. However, this method may have conflicts and interferences between tasks, leading to a decrease in model performance. It is necessary to carefully design the model structure and training strategies to balance the trade-offs between the two tasks.

Finally, the text generation-based ASPE method can flexibly generate complex text structures and handle more flexible and diverse inputs and outputs, performing well in new fields and with fewer samples. However, the training and reasoning of generating models may be more complex and time-consuming.

IV. MABSA EVALUATION CORPUS

Currently, the available datasets for MABSA include Multi-ZOL, Twitter-15, Twitter-17, and MASAD, with

TABLE 2. Statistical information of the Multi-ZOL dataset.

Attributes	Statistic
Number of comments	5528
Number of sentiment labels	10
Aspect-comment Pairs	28469
Average number of aspects in comments	5.45
Average text length in comments	315.11
Minimum text length in comments	5
Maximum text length in comments	8511
Average number of images in comments	4.5
Minimum number of images in comments	1
Maximum number of images in comments	111

TABLE 3. Statistical information for the Twitter-15 and Twitter-17 datasets.

	Twitter-15			Twitter-17		
	Train	Development	Test	Train	Development	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234

Twitter-15 and Twitter-17 datasets being the most commonly used, followed by Multi-ZOL dataset.

A. MULTI-ZOL DATASET

Xu et al. [38] crawled through pages 1-50 of popular mobile reviews on the mobile channel of ZOL.com website. For each phone, only crawl the comments from the top 20 pages. The data crawled includes 114 mobile phone brands and 1318 types of phones. The crawled data contains single modal comments, which are necessary to be filtered out, and as a result, 5288 multimodal comment data are retained. In this dataset, each multimodal comment contains a paragraph of text, an image set, and 1-6 aspects. These six aspects are price-performance ratio, performance configuration, battery life, appearance and feeling, photographing effect, and screen. Pairing each aspect with multimodal comments resulted in a sample of 28469 aspect comment pairs. For each aspect, the comment has an integer sentiment score from 1 to 10, which is used as the sentiment label.

The Multi-ZOL dataset is divided into train, development, and test sets in an 8:1:1 ratio. The number of comment samples with sentiment labels of 7 and 9 in the dataset is 0, so it is set to eight classifications when performing sentiment label classification tasks. The statistical information of the Multi-ZOL dataset is shown in Table 2.

B. TWITTER-15 AND TWITTER-17 DATASETS

Yu et al. [39] selected two publicly available multimodal named entity recognition datasets to construct the Twitter-15 and Twitter-17 datasets, which were collected by Lu et al. [59] and Zhang et al. [60], respectively. These two datasets

include multimodal user posts posted on Twitter from 2014 to 2015 and 2016 to 2017, retaining only posts of person, location, organization, and miscellaneous four entity types, each containing textual content and related image. Due to the fact that these two multimodal datasets only contain manually annotated entities, the author invited three domain experts to annotate each entity sentiment based on text content and associated image. Afterwards, each dataset was randomly divided into three parts in a 3:1:1 ratio: train set, development set, and test set. The statistical information of the Twitter-15 and Twitter-17 datasets is shown in Table 3.

C. MASAD DATASET

Zhou et al. [61] selected 38532 samples from a partial VSO visual dataset (approximately 120000 samples) that can clearly express sentiments and categorized them into seven domains: food, goods, buildings, animal, human, plant, scenery, with a total of 57 predefined aspects. Then crawl the text description of the image and clean each aspect of the data to ensure the high quality of each sample. The MASAD dataset is divided into a train set and a test set, with both positive and negative sentiment polarities. The statistical information of the MASAD dataset is shown in Table 4.

V. MABSA EVALUATION

A. COMMON EVALUATION INDICATORS FOR MABSA

At present, the commonly used evaluation indicators for MABSA include accuracy, precision, recall, and F1 score. The corresponding formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Among them, TP represents the number of correctly predicted positive samples, FP represents the number of incorrectly predicted negative samples, FN represents the number of incorrectly predicted positive samples, and TN represents the number of correctly predicted negative samples. Accuracy refers to the percentage of correctly predicted results in the total sample; Precision refers to the probability of actually being a positive sample among all predicted positive samples; Recall refers to the probability of being predicted as a positive sample among actual positive samples; The F1 score takes into account both precision and recall, achieving the highest level of both and achieving a balance.

B. EVALUATION RESULTS

This section mainly summarizes the evaluation results of existing research methods for MABSA on the Twitter-15, Twitter-17, and Multi-ZOL datasets.

TABLE 4. Statistical information of the MASAD dataset.

	Train			Test			Total		
	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Food	2360	433	2793	592	109	701	2952	542	3494
Goods	2671	1674	4345	743	512	1255	3414	2186	5600
Buildings	1450	970	2420	367	245	612	1817	1215	3032
Animal	3023	2208	5231	1126	670	1796	4149	2878	7027
Human	1999	1838	3837	503	464	967	2502	2302	4804
Plant	2819	2607	5426	1269	947	2216	4088	3554	7642
Scenery	3600	1936	5536	907	490	1397	4507	2426	6933
Total	17922	11666	29588	5507	3437	8944	23429	15103	38532

TABLE 5. Evaluation results of MASC task in the Twitter-15 and Twitter-17 datasets.

Model	Twitter-15		Twitter-17	
	Accuracy	F1	Accuracy	F1
ESAFN [39]	73.38	67.37	67.83	64.22
EF-Net [41]	73.65	67.90	67.77	65.32
HIMT [42]	78.14	73.68	71.14	69.16
HCT [43]	77.62	73.16	68.88	67.45
ITM [44]	78.27	74.19	72.61	71.97
MIGNN [46]	71.98	71.43	64.83	64.42
MVIN [47]	78.60	73.79	73.81	71.02
FGSN [48]	74.69	68.32	69.79	66.90
AMCGC [49]	77.34	72.48	71.15	69.69

TABLE 6. Evaluation results of MASC task in the Multi-ZOL dataset.

Model	Accuracy	F1
MIMN [38]	61.59	60.51
ABAFN [40]	72.88	72.59
HIMT [42]	66.83	66.58
JAAIN [45]	74.57	74.48

Firstly, for the MASC task, the summarized results are shown in Table 5 and Table 6, respectively.

From Table 5, it can be seen that the F1 score of the coarse-to-fine grained ITM network proposed by Yu et al. [44] on the Twitter-15 and Twitter-17 datasets are higher than those of other models. This indicates that in MASC, it is important to first capture image-target relevance, filter out noise caused by irrelevant image, and then align the target with visual area.

From the data in Table 6, it can be seen that the JAAIN model proposed by Zhao et al. [45] performs the best on the Multi-ZOL dataset, indicating that multi-level fusion of aspect information and image and text information, removal of text and image unrelated to a given aspect, and enhancement of the sentiment representation of text modal data in a given aspect can help improve the effectiveness of sentiment classification.

TABLE 7. Evaluation results of ASPE task in the Twitter-15 and Twitter-17 datasets.

Model	Twitter-15			Twitter-17		
	Precision	Recall	F1	Precision	Recall	F1
JML [50]	65.0	63.2	64.1	66.5	65.5	66.0
UMAEC [52]	-	-	58.05	-	-	-
CMMT [53]	64.6	68.7	66.5	67.6	69.4	68.5
DTCA [54]	67.3	69.5	68.4	69.6	71.2	70.4
VLP-MABSA [56]	65.1	68.3	66.6	66.9	69.2	68.0
AoM [57]	67.9	69.3	68.6	68.4	71.0	69.7
GMP [58]	51.67	47.19	49.33	54.28	53.31	53.79

Secondly, for the ASPE task, existing research methods are only based on the Twitter-15 and Twitter-17 datasets, with the evaluation results shown in Table 7.

From Table 7, it can be seen that the DTCA model with cross-modal alignment proposed by Yu et al. [54] generally preforms better, indicating that enhancing the performance of cross attention and achieving text and image alignment is beneficial for ASPE.

VI. POSSIBLE RESEARCH TRENDS FOR MABSA

MABSA is an emerging task in the field of sentiment analysis, and there is currently relatively little research related to it. It can be imagined that there will be more and more research on MABSA in the future, which may include but is not limited to the following aspects:

A. FUSING MORE MODALITIES

Currently, MABSA mainly focuses on the fusion of text and image. How to effectively fuse text and image, solve the mismatch between text and image, and achieve cross-modal alignment is still a research focus. Besides, an intuitive trend of MABSA is to introduce more modalities such as audio and video to be integrated with text and image modalities to obtain richer sentiment information.

B. ENRICHING THE DATASET

Next, we can expect and strive to establish larger and more diverse MABSA datasets, which can contain more text, speech, and image data, and cover more practical application

scenarios, thereby making the MABSA model more accurate and applicable.

C. ACHIEVING MODELS WITH MORE ROBUSTNESS AND INTERPRETABILITY

Models for MABSA are usually complex and require a large amount of data and computational resources for training and inference. One possible future development trend is to improve the robustness and interpretability of models, making them more stable, reliable, and easier to understand. This helps to improve the effectiveness of the model in practical applications and increases user trust and acceptance.

D. MODELING LONG-TERM DEPENDENCY

Sentiment analysis often requires modeling the long-term dependency information of input text, speech, or image. In subsequent research, one challenging trend is more advanced neural network structures or attention mechanism can be studied and applied to better capture and utilize long-term dependencies in multimodal data, thereby improving the performance of sentiment analysis.

E. TRACKING FINE-GRAINED SENTIMENT CHANGES

Sentiment is dynamic that changes over time and context. One interesting trend is to track and analyze fine-grained sentiment changes in multimodal data, such as sentiment transitions and changes in intensity. This helps to better understand users' sentiment states at different time points and contexts, thereby providing more personalized sentiment analysis results.

In summary, MABSA will continue to be developed and improved in the future. With the application of technologies such as fusing more modalities, enriching the dataset, achieving models with more robustness and interpretability, modeling long-term dependency, and tracking fine-grained sentiment changes, we can expect the widespread application and higher accuracy of MABSA in various practical scenarios.

VII. CONCLUSION

In conclusion, this article provides a comprehensive summary of the existing research on MABSA. Firstly, the relevant concepts of MABSA are introduced. Secondly, the existing research methods for MASC and ASPE subtasks are summarized, and the advantages and disadvantages of each type of method are analyzed. Thirdly, the commonly used evaluation indicators and corpus for MABSA are summarized, as well as the evaluation results of existing research methods on the corpus. Finally, the possible research trends for MABSA are envisioned. This paper attempts to establish a relatively complete research view for researchers, hoping to provide some help for further advancing research in this field.

However, it is important to acknowledge the limitations of this article. The literature survey may not have encompassed all the relevant studies on MABSA. In the future,

with the deepening of research, we will strive to expand the scope of literature survey to ensure a more comprehensive coverage of this field. In addition, the future research trend of MABSA may not be detailed enough. Although we have proposed some possible directions, the future research field is very broad, and there are still many unknown challenges and opportunities waiting for further exploration.

REFERENCES

- [1] Z. F. Wang, "Research on multimodal sentiment analysis based on deep learning," M.S. thesis, College Commun. Inf. Eng., Nanjing Univ. Posts Telecommun., Nanjing, China, 2022.
- [2] R. M. Zhao, X. Xiong, S. G. Ju, Z. Z. Li, and C. Xie, "Implicit sentiment analysis for Chinese texts based on a hybrid neural network," *J. Sichuan Univ.*, vol. 57, no. 2, pp. 264–270, 2020, doi: 10.3969/j.issn.0490-6756.2020.02.010.
- [3] L. Yang and M. X. He, "Chinese text sentiment analysis model based on gated mechanism and convolutional neural network," *J. Comput. Appl.*, vol. 41, no. 10, pp. 2842–2848, 2021, doi: 10.11772/j.issn.1001-9081.2020122043.
- [4] Z. Y. Wei, "Research on sentiment analysis of Chinese texts based on BERT," M.S. thesis, College Electron. Eng., Xidian Univ., Xian, China, 2022.
- [5] Q. M. Du, N. Li, W. F. Liu, S. D. Yang, and F. Yue, "Sentiment analysis of Chinese short text combining context and dependent syntactic information," *Com. Sci.*, vol. 50, no. 3, pp. 307–314, 2023, doi: 10.11896/jsjx.211200189.
- [6] K. K. Song, "Research on image sentiment analysis based on deep learning," Ph.D. dissertation, College Inf. Sci. Technol., Univ. Sci. Tech. China, Hefei, China, 2018.
- [7] Y. Q. Miao, Q. Q. Lei, W. Z. Zhang, M. Zhou, and Y. M. Wen, "Research on image sentiment analysis based on multi-visual object fusion," *Appl. Res. Com.*, vol. 38, no. 4, pp. 1250–1255, 2021, doi: 10.19734/j.issn.1001-3695.2020.02.0087.
- [8] J. Y. Yang, "Image emotion analysis combining psychological and deep learning models," Ph.D. dissertation, College Electron. Eng., Xidian Univ., Xian, China, 2022.
- [9] J. N. Geng, "Emotion recognition using user speech," M.S. thesis, College Comput. Sci. Technol., Univ. Sci. Tech. China, Hefei, China, 2021.
- [10] B. W. Cui, "Research on speech emotion analysis algorithm based on pad emotion 3D model," M.S. thesis, College Comput. Sci., Shaanxi Normal Univ., Xian, China, 2022.
- [11] X. Wu, M. T. Hu, and P. Ding, "Multi-modal data representation learning for ceramic coating materials," *J. Shanghai Univ.*, vol. 28, no. 3, pp. 492–503, 2022, doi: 10.12066/j.issn.1007-2861.2383.
- [12] B. Dong, "Research on the representation learning of multimodal data," Ph.D. dissertation, College Comput. Sci., Nat. Univ. Defence Tech., Changsha, China, 2023.
- [13] P. Yu, "Multi-modal fine-grained image classification based on co-attention alignment mechanism," M.S. thesis, College Comput. Sci. Technol., Shandong Univ., Jinan, China, 2022.
- [14] K. Y. Huang, "Research of image-text multimodal representation algorithm based on object-semantics alignment," M.S. thesis, College Comput. Sci. Technol., Huazhong Univ. Sci. and Tech., Wuhan, Hubei, China, 2022.
- [15] W. F. Li, "Research on social emotion classification based on multi-modal fusion," M.S. thesis, College Softw. Eng., Chongqing Univ. Posts Telecommun., Chongqing, China, 2020.
- [16] J. H. Wang, Z. Liu, T. T. Liu, Y. Y. Wang, and Y. J. Cai, "Multimodal sentiment analysis based on multilevel feature fusion attention network," *J. Chin. Inf. Process.*, vol. 36, no. 10, pp. 145–154, 2022.
- [17] Y. Q. Miao, S. Yang, T. L. Liu, W. Z. Zhang, and L. Zhu, "Multimodal sentiment analysis based on cross-modal gating mechanism and improved fusion method," *Appl. Res. Com.*, vol. 40, no. 7, pp. 2025–2030, 2023, doi: 10.19734/j.issn.1001-3695.2022.12.0766.
- [18] R. F. Li, H. Chen, F. X. Feng, Z. Y. Ma, X. J. Wang, and E. Hovy, "Dual graph convolutional networks for aspect-based sentiment analysis," in *Proc. 59th ACL 11th IJCNLP*, 2021, pp. 6319–6329.
- [19] K. Zhang, K. Zhang, M. Zhang, H. Zhao, Q. Liu, W. Wu, and E. Chen, "Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2022, pp. 3599–3610.

- [20] H. Chen, Z. Zhai, F. Feng, R. Li, and X. Wang, "Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2974–2985.
- [21] Y. F. Cheng, J. J. Wu, and F. He, "Aspect level sentiment analysis based on relation gated graph convolutional network," *J. Zhejiang Univ.*, vol. 57, no. 3, pp. 437–445, 2023, doi: [10.3785/j.issn.1008-973X.2023.03.001](https://doi.org/10.3785/j.issn.1008-973X.2023.03.001).
- [22] J. Yu, Q. Zhao, and R. Xia, "Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 1456–1470.
- [23] X. Bao, X. Jiang, Z. Wang, Y. Zhang, and G. Zhou, "Opinion tree parsing for aspect-based sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2023, pp. 7971–7984.
- [24] Y. Zhang and T. R. Li, "Review of comment-oriented aspect-based sentiment analysis," *Comput. Sci.*, vol. 47, no. 6, pp. 194–200, 2020, doi: [10.11896/jsjcx.200200127](https://doi.org/10.11896/jsjcx.200200127).
- [25] L. Wang, H. W. Ma, and H. H. Lv, "Summary of aspect-based sentiment analysis," *J. Comput. Appl.*, vol. 42, no. S2, pp. 1–9, 2022, doi: [10.11772/j.issn.1001-9081.2021122051](https://doi.org/10.11772/j.issn.1001-9081.2021122051).
- [26] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 11, pp. 11019–11038, 2022, doi: [10.1109/TKDE.2022.3230975](https://doi.org/10.1109/TKDE.2022.3230975).
- [27] Y. Li, S. Wang, J. W. Zhu, M. X. Liang, X. Gao, and Z. X. Jiao, "Summarization of aspect-level sentiment analysis," *Comput. Sci.*, vol. 50, no. S1, pp. 34–40, 2023, doi: [10.11896/jsjcx.220400077](https://doi.org/10.11896/jsjcx.220400077).
- [28] Z. Chen, T. Y. Qian, W. L. Li, T. Zhang, S. Zhou, M. Zhong, Y. Y. Zhu, and M. C. Liu, "Low-resource aspect-based sentiment analysis: A survey," *Chin. J. Comput.*, vol. 46, no. 7, pp. 1445–1472, 2023, doi: [10.11897/SPJ.1016.2023.01445](https://doi.org/10.11897/SPJ.1016.2023.01445).
- [29] X. R. Meng, W. Z. Yang, and T. Wang, "Survey of sentiment analysis based on image and text fusion," *J. Comput. Appl.*, vol. 41, no. 2, pp. 307–317, 2021, doi: [10.11772/j.issn.1001-9081.2020060923](https://doi.org/10.11772/j.issn.1001-9081.2020060923).
- [30] J. M. Liu, P. X. Zhang, Y. Liu, W. D. Zhang, and J. Fang, "Summary of multi-modal sentiment analysis technology," *J. Frontiers Comput. Sci. Technol.*, vol. 15, no. 7, pp. 1165–1182, 2021, doi: [10.3778/j.issn.1673-9418.2012075](https://doi.org/10.3778/j.issn.1673-9418.2012075).
- [31] G. W. Chen, P. Z. Zhang, T. Wang, and Q. K. Ye, "Review on multimodal sentiment recognize," *J. Commun. Univ. China*, vol. 29, no. 2, pp. 70–78, 2022, doi: [10.16196/j.cnki.issn.1673-4793.2022.02.009](https://doi.org/10.16196/j.cnki.issn.1673-4793.2022.02.009).
- [32] W. X. Li, H. Y. Mei, and Y. T. Li, "Survey of multimodal sentiment analysis based on deep learning," *J. Liaoning Univ. Tech.*, vol. 42, no. 5, pp. 293–298, 2022, doi: [10.15916/j.issn1674-3261.2022.05.003](https://doi.org/10.15916/j.issn1674-3261.2022.05.003).
- [33] M. Meng, "Sentiment analysis of film criticism based on BERT-TextCNN-B," M.S. thesis, College Math. Phys., Shanghai Normal Univ., Shanghai, China, 2021.
- [34] L. L. Wang, C. L. Yao, X. Li, and X. Q. Yu, "Combining dependency syntactic parsing with interactive attention mechanism for implicit aspect extraction," *Appl. Res. Comput.*, vol. 39, no. 1, pp. 37–42, 2022, doi: [10.19734/j.issn.1001-3695.2021.06.0249](https://doi.org/10.19734/j.issn.1001-3695.2021.06.0249).
- [35] H. S. Chen, J. X. An, Q. H. Tao, and J. Zhou, "Multi-modal sentiment analysis model based on BERT-VGG16," *J. Chengdu Univ. Inf. Tech.*, vol. 37, no. 4, pp. 379–385, 2022, doi: [10.16836/j.cnki.jcuit.2022.04.003](https://doi.org/10.16836/j.cnki.jcuit.2022.04.003).
- [36] S. Zhang, "Research on sentiment analysis technology for multimodal social data," Ph.D. dissertation, College Electron. Inf. Eng., Nanjing Univ. Inf. Sci. Tech., Nanjing, China, 2022.
- [37] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 10, pp. 2048–2057, Jan. 2015, doi: [10.48550/arXiv.1502.03044](https://doi.org/10.48550/arXiv.1502.03044).
- [38] N. Xu, W. J. Mao, and G. D. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI*, 2019, pp. 371–378.
- [39] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 429–439, 2020, doi: [10.1109/TASLP.2019.2957872](https://doi.org/10.1109/TASLP.2019.2957872).
- [40] L. L. Liu, Y. Yang, and J. Wang, "ABAFN: Aspect-based sentiment analysis model for multimodal," *Comput. Eng. Appl.*, vol. 58, no. 10, pp. 193–199, 2022, doi: [10.3778/j.issn.1002-8331.2108-0056](https://doi.org/10.3778/j.issn.1002-8331.2108-0056).
- [41] D. Gu, J. Wang, S. Cai, C. Yang, Z. Song, H. Zhao, L. Xiao, and H. Wang, "Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021, doi: [10.1109/ACCESS.2021.3126782](https://doi.org/10.1109/ACCESS.2021.3126782).
- [42] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1966–1978, Mar. 2022, doi: [10.1109/TAFFC.2022.3171091](https://doi.org/10.1109/TAFFC.2022.3171091).
- [43] K. Chen, X. G. Dong, and X. S. Zhou, "Research on multimodal fine-grained sentiment analysis method based on cross-modal transformer," *Comput. Digit. Eng.*, vol. 50, no. 10, pp. 2270–2275, 2022, doi: [10.3969/j.issn.1672-9722.2022.10.027](https://doi.org/10.3969/j.issn.1672-9722.2022.10.027).
- [44] J. Yu, J. Wang, R. Xia, and J. Li, "Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 4482–4488.
- [45] Y. C. Zhao, S. G. Wang, J. Liao, and D. H. He, "Image-text aspect emotion recognition based on joint aspect attention interaction," *Beijing Univ. Aeronaut. Astronaut.*, vol. 2022, pp. 1–14, Jan. 2022, doi: [10.13700/j.bh.1001-5965.2022.0387](https://doi.org/10.13700/j.bh.1001-5965.2022.0387).
- [46] L. Li and P. Li, "Aspect-level multimodal sentiment analysis based on interaction graph neural network," *Appl. Res. Comput.*, vol. 40, no. 12, pp. 3683–3689, 2023, doi: [10.19734/j.issn.1001-3695.2022.10.0532](https://doi.org/10.19734/j.issn.1001-3695.2022.10.0532).
- [47] X. Y. Wang, W. Q. Pang, and L. J. Zhao, "Multiview interaction learning network for multimodal aspect-level sentiment analysis," *Comput. Eng. Appl.*, vol. 2023, pp. 1–11, Mar. 2023, doi: [10.3778/j.issn.1002-8331.2210-0288](https://doi.org/10.3778/j.issn.1002-8331.2210-0288).
- [48] J. Zhao and F. Yang, "Fusion with GCN and SE-ResNeXt network for aspect based multimodal sentiment analysis," in *Proc. IEEE 6th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, vol. 6, Feb. 2023, pp. 336–340.
- [49] W. Shunjie, C. Guoyong, L. Guangrui, and T. Weibo, "Aspect-level multimodal co-attention graph convolutional sentiment analysis model," *J. Image Graph.*, vol. 28, no. 12, pp. 3838–3854, 2023.
- [50] X. Ju, D. Zhang, R. Xiao, J. Li, S. Li, M. Zhang, and G. Zhou, "Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4395–4405.
- [51] J. M. Dai, W. W. Kong, Z. Wang, and P. Z. Li, "End-to-end aspect-based sentiment analysis model for BERT and LSI," *Comput. Eng. Appl.*, vol. 2023, pp. 1–13, Feb. 2023, doi: [10.3778/j.issn.1002-8331.2303-0220](https://doi.org/10.3778/j.issn.1002-8331.2303-0220).
- [52] R. Zhou, H. Z. Zhu, W. Y. Guo, S. L. Yu, and Y. Zhang, "A unified framework for multimodal aspect-term extraction and aspect-level sentiment classification," *J. Comput. Res. Device*, vol. 60, no. 12, pp. 2877–2889, Mar. 2023, doi: [10.7544/j.issn1000-1239.202220441](https://doi.org/10.7544/j.issn1000-1239.202220441).
- [53] L. Yang, J. C. Na, and J. F. Yu, "Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis," *Inf. Process. Manag.*, vol. 59, no. 5, pp. 1–15, 2022, doi: [10.1016/j.ipm.2022.103038](https://doi.org/10.1016/j.ipm.2022.103038).
- [54] Z. W. Yu, J. Wang, L. C. Yu, and X. J. Zhang, "Dual-encoder transformers with cross-modal alignment for multimodal aspect-based sentiment analysis," in *Proc. 2nd Conf. AACL 12th IJCNLP*, 2022, pp. 414–423.
- [55] W. X. Che, J. Guo, and Y. M. Cui, "Advanced pretrained language model," in *Natural Language Processing: A Pre-trained Model Approach*, 11st ed. Beijing, China: Pub. House Electr. Indu., 2021, ch. 8, sec. 4, pp. 257–260.
- [56] 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.
- [57] R. Zhou, W. Guo, X. Liu, S. Yu, Y. Zhang, and X. Yuan, "AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2023, pp. 8184–8196.
- [58] X. Yang, S. Feng, D. Wang, Q. Sun, W. Wu, Y. Zhang, P. Hong, and S. Poria, "Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2023, pp. 11575–11589.
- [59] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1990–1999.
- [60] Q. Zhang, J. L. Fu, X. Y. Liu, and X. J. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proc. AAAI*, 2018, pp. 5674–5681.

- [61] J. Zhou, J. B. Zhao, J. X. Huang, Q. V. Hu, and L. He, "MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis," *Neurocomputing*, vol. 455, pp. 47–58, Jan. 2021, doi: 10.1016/j.neucom.2021.05.040.



HUA ZHAO received the B.S. degree in computer science and technology from Liaocheng University, China, in 2001, and the M.S. and Ph.D. degrees in computer science and technology from the Harbin Institute of Technology, China, in 2003 and 2008, respectively. She is currently an Associate Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology. Her current research interests include sentiment analysis, natural language processing, and deep learning.



XUEYANG BAI received the B.S. degree in computer science and technology, in 2021. She is currently pursuing the master's degree in electronic information with the Shandong University of Science and Technology, Qingdao, China. Her research interests include natural language processing and named entity recognition.



MANYU YANG received the B.S. degree in computer science and technology, in 2021. She is currently pursuing the master's degree in computer science and technology with the Shandong University of Science and Technology, Qingdao, China. Her research interests include natural language processing and multimodal aspect-based sentiment analysis.



HAN LIU received the B.S. degree in information management and information systems, in 2019. She is currently pursuing the master's degree in library and information with the Shandong University of Science and Technology, Qingdao, China. Her research interests include natural language processing and method entity and relation extraction from scientific and technological literature.

• • •